UNE APPROCHE MÉTHODOLOGIQUE POUR L'ÉLABORATION D'UN SONDAGE APPLIQUÉE À LA POST-STRATIFICATION : LE CAS DE L'ÉTUDE SUR LE TRANSPORT EN COMMUN À ROCK FOREST

par

Jean Cadieux

mémoire présenté à la Faculté des sciences en vue de l'obtention
du grade de maîtrise en mathématiques

FACULTÉ DES SCIENCES
UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, décembre 1996

Sommaire

Ce mémoire présente et discute chacune des étapes de l'élaboration d'un sondage d'opinion et s'attarde notamment aux techniques d'élaboration d'un questionnaire. On y discute de l'importance et de l'utilité de chacune des étapes.

Nous développons aussi toute la théorie entourant le développement des intervalles de confiance pour les plans de sondages simples, stratifiés, de Bernoulli et post-stratifiés. Le but de cette démarche est de mieux comprendre la dynamique intrinsèque des intervalles de confiance issus d'une étude pratique.

Remerciements

Je profite de ce moment privilégié pour remercier tous les gens qui m'ont aidé de près ou de loin tout au long de ma démarche. Parmi ceux-ci mentionnons particulièrement, Bernard Colin, Bernard Courteau, Ernest Monga, Jean-Pierre Samson et Jean Vaillancourt. Tous, à leur manière, ont participé à mon cheminement; ne serait-ce que par une simple discussion philosophique dans les corridors. J'aimerais souligner plus spécialement mon directeur de recherche M. Ernest Monga qui, grâce à sa disponibilité, son écoute et sa confiance, a su allumer en moi un respect et un amour profond pour cette science qu'est la statistique.

J'aimerais aussi remercier ma femme pour ses encouragements et son appui inconditionnel ainsi que ma petite fille qui a fait de nombreux efforts pour me laisser travailler. Il ne faut pas oublier de remercier le comité d'administration de la ville de Rock Forest qui a rendu possible la phase expérimentale de ce mémoire.

Je remercie enfin les professeurs Bernard Colin, Ernest Monga et Alain Boulanger qui ont aimablement accepté de faire partie du jury de ce mémoire.

Table des matières

Liste des tableaux

Liste des figures

Introduction

Les sondages font partie de ces disciplines qui, tout en étant très mal connues dans leur fondement par le grand public, n'en demeurent pas moins abondamment mises en oeuvre dans les aspects les plus divers de la réalité quotidienne. Ce sont les sondages d'opinion et les sondages sur les modes de vie qui, bien adaptés à la médiatisation, constituent la forme la plus connue du domaine des sondages. Cependant, il est faux de restreindre l'emploi des sondages qu'aux sondages d'opinion. En se limitant à quelques domaines familiers, on peut donner des exemples moins connus d'utilisation de techniques de sondage : la vérification de la comptabilité d'entreprise, les contrôles fiscaux effectués par les gouvernements, le contrôle de la qualité de fabrication d'automobiles sur une chaîne de montage, etc.

Malgré tout, c'est la mise en oeuvre d'un sondage d'opinion qui représente pour plusieurs un défi d'une plus grande envergure. En effet, des difficultés majeures comme l'élaboration d'un bon instrument de mesure et le traitement de la non-réponse s'ajoutent aux problèmes de base de la théorie des sondages. Contrairement au traitement de la non-réponse où la littérature abonde, les textes traitant la théorie de l'échantillonnage sont plutôt silencieux, mis à part DESABIE [3], sur l'épineux sujet de la conception de questionnaires. Dans cette optique ce mémoire présente et discute chacune des étapes de l'élaboration d'un sondage d'opinion et s'attarde notamment sur les techniques d'élaboration d'un questionnaire. C'est avec un grand souci de joindre la théorie à la pratique que nous critiquerons toutes les étapes du sondage d'opinion sur le transport en commun qui a été réalisé auprès de la population de la ville de Rock Forest qui a bien voulue se soumettre à l'étude.

Pour couvrir l'ensemble du sujet, sept chapitres sont nécessaires. Le premier chapitre présente l'importante phase de la définition des objectifs d'une enquête. On y discute du rôle capital que jouent les objectifs et les contraintes qu'impose une enquête. Aussi, on met en évidence les différents paramètres de la population à estimer. Finalement, on présente les deux grandes définitions d'une population. Le second chapitre traite des bases de sondage. On y présente les différents types de bases avec leurs avantages et leurs inconvénients. Le troisième chapitre propose un survol des principaux plans d'échantillonnage qui sont utilisés dans les enquêtes. On y justifie le choix du plan de sondage utilisé pour l'étude sur le

transport en commun, le plan stratifié, et on présente toutes les caractéristiques d'usage. Le quatrième chapitre traite de la phase du tirage de l'échantillon. On y présente comment déterminer la taille d'un échantillon afin d'obtenir une précision acceptable. L'algorithme de tirage y est aussi présenté. Le cinquième chapitre de ce mémoire présente l'étape de la conception du questionnaire. On présente et on classe les différents types de questions. On présente aussi quelques conseils de rédaction et quelques grilles d'évaluation pour améliorer un questionnaire. Le sixième chapitre examine les différentes étapes qui suivent la conception et la cueillette des données. En effet, on discute des principales méthodes de collecte de données, de la codification et du traitement de celles-ci et du traitement de la non-réponse. Le septième et dernier chapitre traite de l'analyse des données. On s'attarde notamment à l'aspect déductif de ce type d'analyse.

Chapitre 1

Objectifs d'une enquête

La première étape d'une enquête consiste à évaluer les besoins, actuels et futurs, auxquels pourrait répondre celle-ci. Elle comporte en particulier la description de la population et les caractères à étudier et lorsque l'occasion se présente, la périodicité de l'étude.

Afin de bien couvrir le sujet, ce chapitre se subdivise en trois sections. Dans la première, nous discutons des objectifs et des contraintes de l'enquête. Dans la seconde section, nous allons préciser quels sont les paramètres de la population à estimer au cours de l'enquête sur le transport en commun à Rock Forest. Finalement, dans la dernière section, nous définissons la population à laquelle les résultats de l'étude seront extrapolés.

1.1 Les objectifs et les contraintes de l'enquête

C'est certainement l'étape la plus importante de l'enquête. Négliger cette escale aurait pour conséquence d'élaborer une enquête à l'aveugle, c'est-à-dire sans trop savoir ce que l'on désire. C'est la nature même des objectifs qui déterminera en grande partie, l'orientation ou la philosophie du questionnaire. On y précise ce qu'on veut étudier, quelles sont les vraies valeurs que l'on cherche à estimer, quels seront les champs de l'enquête. Les contraintes sont essentiellement des contraintes de coût et de disponibilité de l'information auxiliaire.

Les attentes de la ville de Rock Forest face aux résultats de cette étude étaient les suivantes: connaître la satisfaction des usagers et mesurer l'efficacité du réseau de transport sur le territoire de la ville. Pour minimiser les coûts et optimiser les résultats, la ville a commandé une seconde étude concernant l'utilité de la bibliothèque. L'espoir des autorités municipales concernant cette seconde partie était d'apporter des réponses aux trois questions suivantes : <<Pourquoi les gens ne vont pas à la bibliothèque?>>, <<Est-ce parce que les services ne sont pas bons?>> et <<Quelle serait l'opinion des gens dans l'éventualité d'un déménagement de la bibliothèque?>>.

Les contraintes relatives à cette enquête ont été essentiellement des contraintes de coût. Par exemple, le fait d'inclure l'étude concernant l'avenir de la bibliothèque a permis de minimiser les coûts de déplacement tout en optimisant la quantité d'information recueillie. L'information auxiliaire nécessaire, qui a permis entre autres de raffiner la stratification, a été fournie gratuitement par le service de taxation de la ville.

1.2 Les paramètres à estimer

L'objectif premier d'une enquête est de fournir des estimations de certains paramètres de la population. Pour une population donnée, ces paramètres sont des constantes. Parmi les paramètres les plus importants, on retrouve le total et la moyenne d'une variable de la population. Aussi, lorsque le caractère à l'étude est mesuré par le présence ou l'absence d'un certain attribut, alors un paramètre d'intérêt est la proportion d'individu de la population qui possèdent l'attribut en question. Si l'on désigne par $X_1, X_2, ..., X_N$ les valeurs prises par une variable X pour les N individus d'une population, on a les définitions ci-après :

le total de la population se note $T_X = \sum_{i=1}^{N} X_i$;

la moyenne de la population se note $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$;

la proportion de la population se note

$$P_y = \frac{\sum_{i=1}^{N} Y_i}{N} \quad \text{où} \quad Y_i = \begin{cases} 1 \text{ si l' individu i possède l' attribut à l' étude.} \\ 0 \text{ sinon} \end{cases}.$$

Dans ce dernier cas, bien que la proportion soit une moyenne particulière, on préfère noter les valeurs de la variable par Y_i parce qu'elles ne prennent que deux valeurs possibles, soient 0 ou 1.

Si nous considérons que les attentes de la municipalité sont les objectifs de l'enquête, on remarque que pour notre étude, les paramètres qui nous intéressent le plus sont les proportions. En effet, nous allons par exemple être tenté d'estimer la proportion de la population qui utilise l'autobus comme moyen de transport, les proportions de familles qui utilisent le transport en commun de façon régulière ou occasionnelle, les proportions des gens

qui sont satisfaits ou non du transport en commun et la proportion des gens qui sont favorables au déménagement de la bibliothèque.

Dans cette enquête, l'estimation est plus importante pour certains paramètres que pour d'autres. En effet, la précision des résultats relative à l'étude du transport en commun est plus importante que celle associée au déménagement éventuel de la bibliothèque. L'influence de cette remarque a d'ailleurs orienté le processus de tirage dans la population. Plus précisément, la sélection de l'échantillon a été effectuée dans le but de favoriser l'estimation des paramètres caractérisant la satisfaction des usagers du transport en commun en fonction de leur proximité relative à un trajet donné.

1.3 Définition de la population

Dans ARDILLY [1], l'auteur considère qu'une population est définie à la base par la conjonction des quatre facteurs suivants: la nature (un individu, un ménage, un logement, une entreprise...), les caractéristiques (sexe, taille du ménage, le type de logement, le secteur d'activité...), la situation géographique (à Rock Forest, dans quel district, sur quel chaîne de montage) et la date. Il y a deux types de définition d'une population. La première est la définition en extension qui prend la forme d'une liste complète des individus; on identifie chaque individu l'un après l'autre de façon suffisamment précise pour qu'il ne subsiste aucune ambiguïté sur les quatre facteurs cités ci-haut. La seconde, dite définition en compréhension, est obtenue par une phrase descriptive qui reprend chacun des facteurs précédents tout en précisant leurs modalités.

Pour cette enquête, la ville de Rock Forest a fourni de l'information concernant l'évaluation foncière de chacun des bâtiments recensés sur le territoire. Compte tenu de la qualité de l'information présente, nous avons opté pour une définition en extension de la population. Ainsi un portrait d'une unité élémentaire de la population est la suivante: sa nature est un ménage, elle a pour caractéristique de payer des taxes à la ville de Rock Forest, elle habite à l'intérieur des limites de la ville et la date de réalisation du projet est l'automne 1995. Selon les données les plus récentes de la Gazette Officielle du Québec du 28 décembre 1994, la population de Rock Forest se chiffrait à 15 119 habitants. Ceux-ci se répartissent dans quelques 5126 logements.

Chapitre 2

Base de sondage

Dès que la population est définie, il faut vérifier s'il existe ou non une liste quelconque de ses individus. Cette liste, communément appelée base de sondage, sert de support lors du tirage de l'échantillon. On recherche la base de sondage la mieux adaptée à la situation en tenant compte de sa qualité et des coûts encourus. On peut éventuellement utiliser plusieurs bases de données provenant de diverses sources afin de fabriquer la base de sondage adéquate. Mais parfois, pour une raison quelconque, la base recherchée n'est que partielle ou pas disponible du tout. La présence ou l'absence de cette base de sondage ainsi que la qualité de l'information qu'elle contient aura une conséquence directe sur la nature du plan d'échantillonnage qu'on utilisera.

Pour donner une meilleure vision de l'étendue du sujet, ce chapitre se divise en trois sections. Dans la première section nous présentons les différentes définitions et propriétés théoriques souhaitables d'une base de sondage. Dans la seconde, nous énonçons les différents types d'erreurs qui se retrouvent dans une base de sondage. Finalement, nous examinons et critiquons la base de sondage qui a été utilisée pour l'étude.

2.1 Définitions et propriétés d'une base de sondage

Il existe deux types de base de sondage. Le premier type est une base dite de liste. Elle est constituée par une liste d'identifiants de chacun des individus de la population à l'étude. Ce type de base matérialise en fait la définition exhaustive de la population. Le deuxième type est une base dite aréolaire qui est constituée par des aires géographiques bien délimitées. Ce type de base s'associe à la définition en compréhension de la population puisqu'alors, dans la majorité des cas, les individus appartenant à une aire seront décrits à l'aide d'une phrase. En somme, on cherche à ce que tous les éléments de la population détiennent une chance, c'est-à-dire une probabilité, si possible connue et contrôlée, de figurer dans l'échantillon.

Une base de sondage doit posséder trois propriétés fondamentales. Premièrement, une base doit repérer chacune des unités de la population sans ambiguïté. Il est impératif qu'on ne prenne pas un individu pour un autre. En effet, l'absence de cette propriété modifie la probabilité d'inclusion des individus, ce qui conduit inévitablement à des biais lors des estimations. Deuxièmement, une base doit être exhaustive. Cela signifie qu'on ne doit oublier personne. Si ce n'est pas le cas, on dit que la base de sondage est incomplète ou encore qu'elle présente des défauts de couverture. Finalement, une base de sondage doit être sans double compte. C'est-à-dire qu'aucun individu ne doit être présent deux fois dans la base, même sous deux identifiants différents.

Aux trois propriétés précédentes peut être ajoutée une dernière qui, sans être indispensable, peut rehausser la qualité de l'information contenue dans la base. Une base de sondage doit être munie d'une information auxiliaire de bonne qualité apportée par des variables disponibles à l'enquêteur. Cette information peut s'avérer utile pour améliorer la méthode de tirage, la qualité des estimateurs ou même les deux.

Dans ARDILLY [1], l'auteur souligne qu'il est extrêmement difficile en pratique de s'affranchir complètement des manques aux trois grandes propriétés précédentes. L'important étant de juger de leur impact et de ne conserver que les bases faiblement imparfaites qui sont les mieux adaptées.

2.2 Les types d'erreurs que peut contenir une base de sondage

Différents types d'imperfections sont susceptibles de se trouver dans une base. Parmi les types les plus répandus, on retrouve les erreurs d'échantillonnage, les erreurs d'observation (volontaires ou non) et les erreurs dues au défaut de couverture. Il est important de comprendre et surtout d'être conscient qu'il est pratiquement impossible de contrôler ou de mesurer les erreurs contenues dans une base de sondage. Sur ce sujet, ARDILLY [1] précise qu'il est inutile de raffiner une méthode de tirage si les erreurs d'observation et de défaut de couverture sont importantes.

L'erreur d'échantillonnage est reliée au fait qu'on ne dispose que des observations d'un échantillon S sur lequel on base toutes nos statistiques et non pas sur toute la population. On peut assimiler à des erreurs d'échantillonnage certaines erreurs induites par les imperfections

de la base de sondage. En effet, l'exhaustivité partielle, les doubles comptes non percevables ainsi que le vieillissement de l'information contribuent à augmenter le biais des estimateurs. On comprend que seul un recensement dépourvu de données manquantes possède théoriquement une erreur d'échantillonnage nulle.

L'erreur de couverture est un second type d'erreur. Elle est causée par une base de sondage incomplète. La population n'est donc pas correctement couverte par la base. Dans ce cas, il y a, dès l'origine, un biais de l'estimation qu'on ne peut pas mesurer. Pour bien illustrer l'aspect corrosif de ce phénomène, il est possible que les unités manquantes dans la base, qui ont une probabilité nulle d'être sélectionnées, ne soient pas bien représentées dans l'échantillon. Donc l'extrapolation des résultats à l'ensemble de la population peut s'en trouver biaisée.

Finalement, l'erreur d'observation caractérise le dernier type d'erreur qu'on peut retrouver dans une base. Elle tient compte de l'inexactitude de l'information présente dans la base, soit parce que l'information a été recueillie de façon inexacte, soit qu'elle a été mal transcrite ou codée, ... Contrairement à l'erreur de couverture où l'information est absente, l'information est ici présente mais différente, voire même très différente, de la vraie valeur.

2.3 Base des taxes municipales: la base la mieux adaptée

Dans notre enquête, la base de sondage a été fournie par la ville. Cette base contenait la liste de toutes les évaluations foncières pour le territoire de Rock Forest. Les informations présentes dans cette base, pour chacun des bâtiments du territoire, étaient les suivantes: adresse civique complète, code représentant la fonctionnalité du bâtiment (commerce, résidence, église, ...), le nombre de logements, un code représentant l'appréciation globale de la propriété (qualité de l'entretien, apparence, ...), l'évaluation de la maison, la superficie du terrain, l'évaluation du terrain et finalement l'évaluation foncière qui est la somme des valeurs de la maison et du terrain. On notera que le territoire compte exactement 5796 logements habitables par des ménages. Cette base permet donc de repérer un ménage sans aucune ambiguïté. Par sa forme elle correspond à la définition qu'on a donnée d'une base de liste.

On ne peut passer sous silence la qualité de l'information auxiliaire contenue dans l'évaluation foncière de chacune des propriétés. Généralement, cette information est

fortement liée au revenu des ménages. En effet, un ménage avec un faible revenu ne peut, le plus souvent, se payer une propriété de luxe. Nous nous sommes d'ailleurs servis de cette information pour améliorer la méthode de tirage. En effet, nous avions raison de croire que les ménages dont la valeur de la maison dépassait un certain montant, par exemple plus de 200 000$, étaient moins sujets à prendre l'autobus.

Différentes erreurs d'échantillonnage incontrôlables sont naturellement reliées à ce type de fichier: le vieillissement de l'information et le double comptage. D'une part, les évaluateurs de la ville procèdent au rafraîchissement de toutes les évaluations foncières des maisons existantes à l'intérieur d'une période de quatre ans; d'autre part, il est possible qu'un ménage possède une maison et un chalet sur le même territoire. Comme l'évaluation foncière des nouveaux bâtiments s'effectue à l'intérieur d'une année, on remarque que les résidences nouvellement construites ne sont pas toutes incluses au fichier. Ce type d'imperfection, éventuellement peu fréquente, introduit tout de même des erreurs de couverture.

Mentionnons aussi que quelques erreurs d'observation ont été décelées a posteriori. Certains bâtiments détruits ou désaffectés depuis peu sont encore inscrits sur la liste comme étant des logements habités par des ménages. On peut aussi critiquer la qualité de l'information auxiliaire en considérant que l'évaluation foncière renferme une large part de subjectivité. Il est fort probable que les valeurs de l'évaluation foncière ne soient pas similaires d'un bâtiment à l'autre ou d'un évaluateur à l'autre.

De ces constatations, nous sommes en mesure de nous prononcer sur la qualité globale de cette base. Premièrement elle possède la propriété de repérer chacune des unités de la population sans ambiguïté. Deuxièmement, malgré les nouvelles constructions, elle est assez exhaustive. Troisièmement, elle semble assez bien résister aux doubles comptes. Malgré les imperfections que cette base contient, on constate en vertu de la qualité de l'information présente, qu'elle constitue une bonne base de sondage.

Chapitre 3

Plan d'échantillonnage

Le tirage de l'échantillon dépend du plan d'échantillonnage mis en oeuvre. Un plan d'échantillonnage définit la méthode de tirage et l'expression des estimateurs des différents paramètres de la population que nous désirons évaluer. La méthode de tirage est le processus choisi pour tirer un échantillon. Nous allons nous intéresser aux probabilités d'inclusion de chaque individu de la population dans l'échantillon lui-même. L'expression de l'estimateur, quant à lui, est la formule choisie pour estimer le paramètre inconnu qui nous intéresse.

Il existe deux classes de plan d'échantillonnage: la classe des plans probabilistes et la classe des plans empiriques. Par définition, les plans probabilistes se caractérisent par le fait que chaque individu a une probabilité connue à l'avance d'appartenir à l'échantillon. Parmi ceux-ci, on retrouve le plan aléatoire simple et le plan stratifié. Par opposition, les plans d'échantillonnage empiriques se caractérisent par le fait qu'on ne peut calculer les dites probabilités d'inclusion. Parmi ceux-ci on retrouve les méthodes des quotas et des unités types. Cependant, compte tenu de l'information présente dans la base de sondage, nous sommes en mesure d'élaborer un plan d'échantillonnage probabiliste.

Dans un premier temps nous allons présenter le plan d'échantillonnage aléatoire simple. Nous allons y dégager l'estimateur du paramètre de la proportion qui nous intéresse. Ensuite, dans un deuxième temps, nous allons présenter le plan de sondage stratifié. Là aussi nous allons y dégager l'estimateur de la proportion. Dans un troisième et dernier temps, à la lumière des deux sections précédentes, nous présentons et justifions notre choix du plan d'échantillonnage.

Dans tous les développements qui suivent, le nombre total d'unités élémentaires sera dénoté N. Les caractères à l'étude seront des variables aléatoires que nous dénoterons par les lettres X, Y, etc. La valeur du caractère X pour la i-ème unité élémentaire de la population sera désignée par x_i. Lorsque nous serons en présence d'un échantillon de taille n issu d'une

population de N éléments, nous dénoterons par $x_1, x_2, \ldots, x_n$ les valeurs du caractère X observées dans l'échantillon.

3.1 Le plan aléatoire simple (P.A.S.)

Dans cette section nous présentons les fondements théoriques du plan d'échantillonnage aléatoire simple. Cette méthode de tirage revêt une importance historique puisqu'elle a permis d'établir la base de la théorie de l'échantillonnage. En bref, cette méthode consiste simplement à affecter à chacun des échantillons possibles, de taille n fixée à l'avance, la même probabilité d'être sélectionné.

Afin de bien couvrir chacune des facettes de ce plan, nous allons tout d'abord présenter une description du plan aléatoire simple, nous enchaînerons ensuite avec l'expression des estimateurs de proportions ainsi que l'expression d'un intervalle de confiance de niveau $1-\alpha$.

3.1.1 Description du plan aléatoire simple

Le sondage aléatoire simple consiste à tirer dans la population de taille N un échantillon de taille fixée n sans remise, sans aucune manipulation préalable de la population et de façon à ce que chaque individu ait la même probabilité d'inclusion.

THÉORÈME 3.1.1

Dans un plan aléatoire simple, la probabilité d'inclusion d'un individu dans l'échantillon est de $\frac{n}{N}$.

Démonstration :

En effet, soit X_i, la valeur du caractère d'un élément quelconque de la population. Il y a $\binom{N-1}{n}$ échantillons de taille n qui ne contiennent pas l'élément en question.

Donc la probabilité pour que l'élément ne soit pas retenu dans l'échantillon est

$$\binom{N-1}{n} \Big/ \binom{N}{n} = 1 - \frac{n}{N}$$

et donc $\frac{n}{N}$ qu'il soit sélectionné.

■

3.1.2 Expression de l'estimateur de la proportion pour le P.A.S.

Pour le plan aléatoire simple, l'estimateur sans biais de la proportion P_y de la population est le suivant :

$$p_y = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ où } y_i = \begin{cases} 1 \text{ si l'individu i possède le caractère} \\ 0 \text{ sinon} \end{cases}.$$

On remarque que cet estimateur est naturel et d'une certaine façon légitime, puisqu'il représente tout simplement la proportion des individus de l'échantillon qui possèdent la caractéristique étudiée.

THÉORÈME 3.1.2

L'estimateur de la proportion

$$p_y = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ où } y_i = \begin{cases} 1 \text{ si l'individu i possède le caractère} \\ 0 \text{ sinon} \end{cases}$$

est sans biais. De plus, sa variance est donnée par l'expression suivante :

$$Var(p_y) = \frac{P_y(1-P_y)}{n}\left(\frac{N-n}{N-1}\right) \text{ pour } n \geq 1.$$

Démonstration :

Montrons premièrement que cet estimateur pour la proportion est sans biais. Il suffit de montrer que $E(p_y) = P_y$. Définissons l'indicateur suivant :

$$I_i = \begin{cases} 1 \text{ si l'individu i est dans l'échantillon} \\ 0 \text{ sinon} \end{cases}.$$

Alors sous le plan aléatoire simple, on a $\mathrm{P}(I_i=1)=\frac{n}{N}\ \forall i=1,2,\ldots,N$ et I_i suit une loi de Bernoulli de paramètre $\frac{n}{N}$. Alors on a

$$\begin{aligned} Var(I_i) &= E\left(I_i^2\right)-\left[E(I_i)\right]^2 \\ &= E\left(I_i\right)-\left[E(I_i)\right]^2 \\ &= \frac{n}{N}-\left[\frac{n}{N}\right]^2 \\ &= \frac{n}{N}\left(1-\frac{\mathrm{n}}{\mathrm{N}}\right)\ \forall i=1,2,\ldots,N \end{aligned}$$

puisque $I_i=I_i^2$ et que

$$\mathrm{E}(I_i)=\mathrm{P}(I_i=1)=\frac{n}{N}\ \forall i=1,2,\ldots,N.$$

Et comme on peut écrire

$$p_y=\frac{1}{n}\sum_{i=1}^{n}y_i=\frac{1}{n}\sum_{i=1}^{N}I_iY_i\ ,$$

on obtient finalement que

$$\mathrm{E}\left(p_y\right)=\frac{1}{n}\sum_{i=1}^{N}Y_i\mathrm{E}(I_i)=\frac{1}{n}\sum_{i=1}^{N}Y_i\left(\frac{n}{N}\right)=\left(\frac{1}{N}\right)\sum_{i=1}^{N}Y_i=P_y.$$

Deuxièmement, le calcul de la variance s'effectue de la façon suivante :

$$Var\left(p_y\right)=Var\left(\frac{1}{n}\sum_{i=1}^{N}I_iY_i\right)=\frac{1}{n^2}Var\left(\sum_{i=1}^{N}I_iY_i\right)=\left[\sum_{i=1}^{N}Y_i^2Var(I_i)+\sum_{i=1\neq j}^{N}\sum_{j=1}^{N}Y_iY_j\,\mathrm{cov}\left(I_i,I_j\right)\right].$$

Soit S un échantillon, on remarque que $I_iI_j=I_{ij}=\begin{cases}1 & \text{si i et j sont dans S}\\ 0 & \text{sinon}\end{cases}$. D'où

$$\mathrm{P}(I_{ij}=1)=\frac{n(n-1)}{N(N-1)}\ \forall i=1,2,\ldots,N.$$

Ainsi on obtient que

$$\operatorname{cov}(I_i, I_j) = \mathrm{E}(I_i I_j) - \mathrm{E}(I_i)\mathrm{E}(I_j) = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2.$$

La variance s'écrit donc :

$$\begin{aligned}
Var(p_y) &= \frac{1}{n^2}\left[\frac{n}{N}\left(1-\frac{n}{N}\right)\sum_{i=1}^{N} Y_i^2 + \left[\frac{n}{N}\left(\frac{n-1}{N-1}\right) - \left(\frac{n}{N}\right)^2\right]\Sigma_{i\neq j}\Sigma Y_i Y_j\right] \\
&= \frac{1}{n^2}\left[\frac{n}{N}\left(\frac{N-n}{N}\right)\left(\frac{N-1}{N-1}\right)\sum_{i=1}^{N} Y_i^2 + \frac{n}{N}\left[\left(\frac{N(n-1)-(N-1)n}{N(N-1)}\right)\right]\Sigma_{i\neq j}\Sigma Y_i Y_j\right] \\
&= \frac{1}{n^2}\left[\frac{n}{N}\left(\frac{N-n}{N-1}\right)\left(\frac{N-1}{N}\right)\sum_{i=1}^{N} Y_i^2 - \frac{n}{N}\left(\frac{N-n}{N-1}\right)\left[\left(\frac{1}{N}\right)\right]\Sigma_{i\neq j}\Sigma Y_i Y_j\right] \\
&= \frac{1}{n^2} n\left(\frac{N-n}{N-1}\right)\left[\left(\frac{N-1}{N^2}\right)\sum_{i=1}^{N} Y_i^2 - \left(\frac{1}{N^2}\right)\Sigma_{i\neq j}\Sigma Y_i Y_j\right] \\
&= \frac{1}{n}\left(\frac{N-n}{N-1}\right)\left[\left(\frac{1}{N}\right)\sum_{i=1}^{N} Y_i^2 - \left(\frac{1}{N^2}\right)\sum_{i=1}^{N} Y_i^2 - \left(\frac{1}{N^2}\right)\Sigma_{i\neq j}\Sigma Y_i Y_j\right] \\
&= \frac{1}{n}\left(\frac{N-n}{N-1}\right)\left[\left(\frac{1}{N}\right)\sum_{i=1}^{N} Y_i^2 - \left(\frac{1}{N^2}\right)\left[\sum_{i=1}^{N} Y_i^2 + \Sigma_{i\neq j}\Sigma Y_i Y_j\right]\right] \\
&= \frac{1}{n}\left(\frac{N-n}{N-1}\right)\left[\left(\frac{1}{N}\right)\sum_{i=1}^{N} Y_i^2 - \left(\frac{1}{N^2}\right)\left[\sum_{i=1}^{N} Y_i\right]^2\right] \text{ mais } Y_i^2 = Y_i \\
&= \frac{1}{n}\left(\frac{N-n}{N-1}\right)\left[\left(\frac{1}{N}\right)\sum_{i=1}^{N} Y_i - \left[\left(\frac{1}{N}\right)\sum_{i=1}^{N} Y_i\right]\left[\left(\frac{1}{N}\right)\sum_{i=1}^{N} Y_i\right]\right] \text{ et comme } P_y = \left(\frac{1}{N}\right)\sum_{i=1}^{N} Y_i \\
&= \frac{1}{n}\left(\frac{N-n}{N-1}\right)\left[P_y - P_y P_y\right] \\
&= \frac{P_y(1-P_y)}{n}\left(\frac{N-n}{N-1}\right)
\end{aligned}$$

■

Connaissant la variance de l'estimateur de la proportion P_y, nous sommes en mesure de construire un intervalle de confiance approximatif de niveau $1-\alpha$. Pour une taille d'échantillon supérieure à deux éléments, nous obtenons l'intervalle de confiance désiré :

$$p_y \pm z_{\alpha/2}\sqrt{\hat{V}ar(p_y)},$$

où $z_{\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ d'une loi normale centrée et réduite et où l'estimateur sans biais de la variance théorique est donné par

$$\hat{V}ar(p_y) = \frac{p_y(1-p_y)}{n-1}\left(\frac{N-n}{N}\right),$$

comme le montre le résultat ci-dessous.

THÉORÈME 3.1.3

Sous le plan aléatoire simple, l'estimateur de $Var(p_y)$ donné par

$$\hat{V}ar(p_y) = \frac{p_y(1-p_y)}{n-1}\left(\frac{N-n}{N}\right) \text{ pour } n \geq 2$$

est sans biais.

Démonstration :

Il s'agit de montrer que $E\left[\hat{V}ar(p_y)\right] = Var(p_y)$. Pour ce faire, calculons $E\left[p_y^2\right]$. On a :

$$E\left[p_y^2\right] = Var\left[p_y\right] + \left(E\left[p_y\right]\right)^2 = \frac{P_y(1-P_y)}{n}\left(\frac{N-n}{N-1}\right) + P_y^2;$$

donc

$$\begin{aligned}
E\left[\hat{V}ar(p_y)\right] &= \left(\frac{N-n}{N(n-1)}\right)\left(E\left[p_y\right] - E\left[p_y^2\right]\right) \\
&= \left(\frac{N-n}{(n-1)N}\right)\left(P_y - P_y^2 - \frac{P_y(1-P_y)}{n}\left(\frac{N-n}{N-1}\right)\right) \\
&= P_y(1-P_y)\left(\frac{N-n}{(n-1)N}\right)\left(1 - \left(\frac{N-n}{n(N-1)}\right)\right) \\
&= P_y(1-P_y)\left(\frac{N-n}{(n-1)N}\right)\left(\frac{nN-n-N+n}{n(N-1)}\right) \\
&= P_y(1-P_y)\left(\frac{1}{n}\right)\left(\frac{N-n}{N-1}\right) \\
&= Var(p_y)
\end{aligned}$$

■

3.2 Le plan stratifié

L'échantillonnage simple sélectionne au hasard un nombre n d'individus de la population. Il peut arriver que cet échantillon ne soit pas du tout représentatif de la population. Pour remédier à cet état de choses, on peut utiliser la stratification. La stratification consiste à partitionner la population en sous-groupes distincts, appelés strates, de sorte qu'un élément (ou encore une unité) n'appartienne qu'à une et une seule strate. Le critère de stratification adopté est celui qui rend la strate la plus homogène possible par rapport à l'ensemble des variables étudiées. Dans ce plan, l'échantillonnage est fait séparément dans chaque strate. Si le processus de tirage dans chaque strate correspond à celui du plan aléatoire simple, nous disons que nous avons un plan aléatoire stratifié, que nous appelerons tout simplement plan stratifié par la suite.

Dans cette section, nous présenterons tout d'abord une description du plan stratifié. Nous donnerons ensuite l'expression de l'estimateur de la proportion de la population, et nous énonçerons et démontrerons quelques résultats concernant cet estimateur.

3.2.1 Description du plan stratifié

En général, la stratification est utilisée pour des besoins de représentativité, de gain de précision et aussi lorsque l'on veut dans une enquête globale faire des inférences particulières sur des sous-groupes de la population. En d'autres mots, cette méthode permet d'étudier l'ensemble d'une population et chacune des différentes sous-populations qui la composent.

Comme le montre la figure 3.2.1 suivante, un plan stratifié est donc un plan pour lequel la population est subdivisée en L strates mutuellement exclusives. La taille de la population de la strate h est définie par N_h et la taille totale de la population est donc donnée par

$$N = \sum_{h=1}^{L} N_h .$$

Suivant la façon dont on répartit notre échantillon de taille n à travers des L strates, un échantillon de taille n_h, connue à l'avance, est tiré indépendamment dans chacune des strates.

Dans la strate h, $1 \le h \le L$, on note par P_{hy} la proportion des individus de cette sous-population possédant une certaine caractéristique. De même, on note par p_{hy} la proportion des individus de l'échantillon possédant cette même caractéristique.

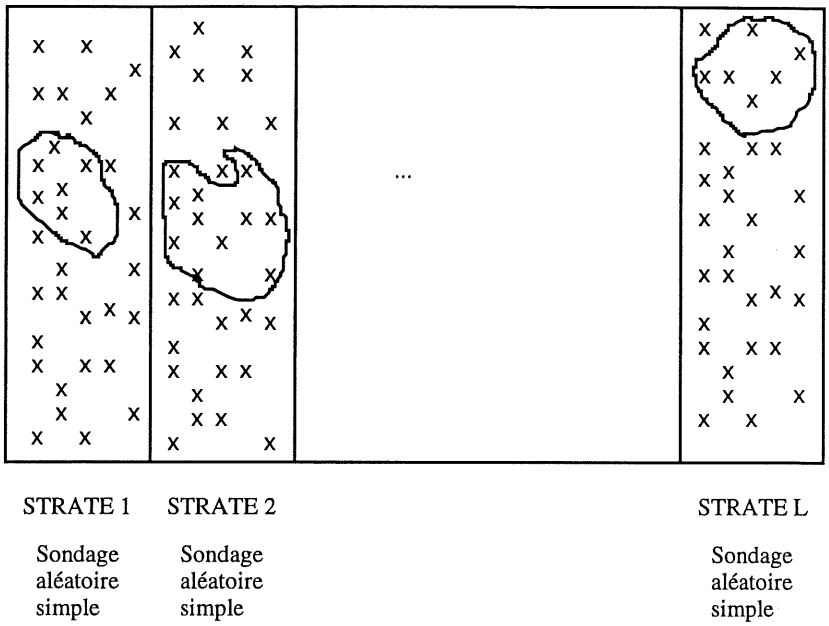

Figure 3.2.1: La stratification d'une population

3.2.2 Expression de l'estimateur de la proportion pour le plan stratifié

THÉORÈME 3.2.1

Soit p_{hy} la proportion empirique des observations issue de la strate h, $1 \le h \le L$, possédant un quelconque caractère prédéfini. Dans le plan stratifié, l'estimateur de la proportion de la population

$$p_{y,str} = \frac{1}{N} \sum_{h=1}^{L} N_h p_{hy}$$

est sans biais. De plus sa variance théorique s'exprime de la façon suivante :

$$Var\left(p_{y,str}\right) = \frac{1}{N^2} \sum_{h=1}^{L} N_h^2 \left[\frac{P_{hy}\left(1-P_{hy}\right)}{n_h}\right]\left(\frac{N_h - n_h}{N_h - 1}\right) \quad \text{Où} \quad p_{hy} = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} \quad \text{et} \quad P_{hy} = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}.$$

Démonstration :

Montrons que cet estimateur proposé pour la proportion est sans biais. Il suffit de montrer que $E(p_{y,str}) = P_y$. Pour effectuer cette démonstration, nous allons utiliser le fait que le tirage dans chacune des strates s'effectue indépendamment selon un plan aléatoire simple. On peut alors écrire que

$$E\left[p_{y,str}\right] = E\left[\frac{1}{N}\sum_{h=1}^{L} N_h p_{hy}\right] \text{ où } p_{hy} = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$$

p_{hy} est la proportion échantillonnale de la strate h dans laquelle nous avons effectué un tirage aléatoire simple. En d' autres mots, on a : $E\left[p_{hy}\right] = P_{hy}$

$$= \frac{1}{N}\sum_{h=1}^{L} N_h E\left[p_{hy}\right]$$

$$= \frac{1}{N}\sum_{h=1}^{L} N_h P_{hy}$$

$$= P_y.$$

De même, afin de trouver l'expression de la variance théorique, nous allons encore profiter du fait que le tirage dans chacune des strates est fait indépendamment à l'aide du plan aléatoire simple :

$$Var\left[p_{y,str}\right] = Var\left[\sum_{h=1}^{L}\frac{N_h}{N}p_{hy}\right] \text{ posons } W_h = \frac{N_h}{N}, \text{ on a alors}$$

$$= Var\left[\sum_{h=1}^{L} W_h p_{hy}\right]$$

$$= \sum_{h=1}^{L} W_h^2 Var\left[p_{hy}\right] + \sum_{h=1}^{L}\sum_{\substack{h'=1 \\ h\neq h'}}^{L} W_h W_{h'} \operatorname{cov}\left[p_{hy}, p_{h'y}\right]$$

Mais les tirages sont indépendants d' une strate à l' autre. Donc $\operatorname{cov}\left[p_{hy}, p_{h'y}\right] = 0$. De plus, $Var\left[p_{hy}\right]$ est la variance d' un estimateur de la proportion dans la strate h dans laquelle on a effectué un tirage aléatoire simple.

$$= \sum_{h=1}^{L} W_h^2 \left[\frac{P_{hy}\left(1-P_{hy}\right)}{n_h}\left(\frac{N_h - n_h}{N_h - 1}\right)\right]$$

$$= \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 \left[\frac{P_{hy}\left(1-P_{hy}\right)}{n_h}\right]\left(\frac{N_h - n_h}{N_h - 1}\right)$$

■

Connaissant la variance théorique de notre estimateur de la proportion, nous obtenons finalement l'intervalle de confiance en utilisant la forme suivante :

$$p_{y,str} \pm z_{\alpha/2}\sqrt{\hat{V}ar\left(p_{y,str}\right)},$$

où l'estimateur de la variance théorique est donné par :

$$\hat{V}ar\left(p_{y,str}\right) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 \left[\frac{p_{hy}\left(1-p_{hy}\right)}{n_h - 1}\right]\left(\frac{N_h - n_h}{N_h}\right).$$

Comme pour le plan aléatoire simple, cet estimateur de la variance est un estimateur sans biais de $Var\left[p_{y,str}\right]$. Pour le voir, il suffit d'appliquer les résultats du théorème 3.1.3 dans chaque strate et d'utiliser l'indépendance des strates.

3.2.3 Les différents types d'affectation

Une fois les L strates choisies, la façon dont l'enquêteur distribue la taille d'échantillon globale à chacune des strates, s'appelle l'affectation. Il existe plusieurs types d'affectation. Parmi les plus courantes on retrouve l'affectation proportionnelle, l'affectation optimale et l'affectation identique.

L'affectation proportionnelle répartit la taille de l'échantillon globale selon le poids de chacune des strates. En fait, ce type d'affectation consiste à prendre la même fraction d'échantillonnage n_h/N_h dans chacune des strates. Cette quantité est alors égale au taux d'échantillonnage n/N. C'est de cette façon que l'enquêteur est capable de calculer la fraction d'échantillon qu'il doit tirer dans chacune des strates : $n_h = N_h\left(n/N\right)$.

L'affectation optimale quant à elle, fixe les tailles d'échantillon dans chacune des strates en optimisant la précision des résultats tout en respectant la contrainte du budget. On y arrive en minimisant l'expression de la variance $\text{var}\left[\hat{P}_{y,str}\right]$ sous la contrainte du budget. Le résultat de l'optimisation montre qu'il faut mettre des efforts particuliers dans les strates dont la population est importante, dans les strates où la diversité est la plus forte et dans les strates où le coût associé à la collecte des données est faible. Dans le cas où le coût par observation est

le même dans chacune des strates, l'affectation optimale prend le nom d'affectation de Neyman. On trouvera un traitement satisfaisant sur la notion d'affectation optimale dans COCHRAN [2].

Un autre type d'affectation consiste à allouer le même nombre d'éléments à chaque strate; $n_h = n/L \ \forall h = 1,2,...,L$. On applique en pratique ce type d'affectation quand on a des raisons de croire que les variances intrastrates sont égales. Elle peut être aussi utilisée lorsqu'on a besoin d'une sur-représentativité chez un petit groupe de la population qu'on croit différent.

3.3 Choix du plan d'échantillonnage

Le territoire de Rock Forest est divisé en trois zones très distinctes : urbaine, riveraine et rurale. Nous soupçonnions que les besoins des familles, en terme de transport en commun, pouvaient être différents d'une zone d'habitation à l'autre, nous avons donc élaboré un plan d'échantillonnage de type stratifié.
Au moment de l'étude les populations des trois strates étaient les suivantes: la strate urbaine contenait exactement 4086 logements habitables et était nettement la plus grande des trois, la strate riveraine contenait exactement 982 logements habitables et finalement, la strate rurale contenait exactement 728 logements habitables.

Le choix de l'affectation a été plus délicat . Elle a été choisie en tenant compte des deux points suivants. D'une part nous avions des raisons de penser que la population constituant la strate riveraine était de type "nomade", dans le sens où les propriétaires ou locataires n'habitent ces logements qu'en saison estivale. D'autre part, la variabilité de la valeur des maisons de la population constituant la strate rurale était la plus grande des trois strates. Pour ces raisons, le type d'allocation que nous avons utilisé pour l'étude correspond à celui de l'affectation identique. Cette stratégie nous a permis de nous assurer une plus grande représentativité là où les sujets nous semblaient les plus différents. La taille de l'échantillon sera discutée dans le prochain chapitre.

Chapitre 4

Le tirage de l'échantillon

Rappelons qu'un plan d'échantillonnage définit la méthode de tirage ainsi que l'expression des estimateurs des paramètres de la population que nous désirons évaluer. Pour les raisons que nous avons présentées dans le chapitre précédent, nous avons opté pour le plan d'échantillonnage de type stratifié jumelé à une affectation identique. Rappelons également que le tirage de l'échantillon dans chacune des strates est effectué selon un plan aléatoire simple. Dans ce chapitre, nous allons faire une analyse minutieuse du tirage de l'échantillon .

Pour bien couvrir le sujet, nous proposons de diviser ce chapitre en deux sections. Dans la première nous présentons et justifions la taille de notre échantillon global. Dans la seconde section, nous présentons l'algorithme de tirage que nous avons utilisé pour sélectionner nos trois échantillons.

4.1 Taille de l'échantillon

Le nombre de logements habitables dans les limites de la ville de Rock Forest est de 5796. Afin de mieux identifier notre population cible, nous avons redéfini la population de base en enlevant les ménages que nous soupçonnions indépendants du transport en commun. Plus précisément, nous avions des raisons de croire qu'un ménage qui possède une maison évaluée à plus de 165 000$ ressent moins le besoin de prendre l'autobus comme moyen de transport. Il y avait dans la population de base 175 ménages de ce genre. De même, nous avions des raisons de croire qu'un ménage qui possède une maison dont la valeur est inférieure à 25 000$ ne ressentait pas le besoin de prendre l'autobus comme moyen de transport. En fait, il s'agissait bien souvent de petits chalets sur le bord de l'eau ou de maisons désaffectées. Il y avait dans notre population initiale 495 de ces bâtiments. Ainsi, notre population susceptible de prendre le transport en commun était distribuée dans les 5126 logements habitables restants.

Si nous observons la variance théorique de notre estimateur de la proportion dans le plan stratifié, nous pouvons conclure que plus la taille de l'échantillon est grande, plus notre sondage est précis. Cependant, nous ne disposions ni d'un budget et ni d'un temps infini. Ainsi, pour trouver la taille d'échantillon, nous avons dû nous conformer aux contraintes suivantes: la disponibilité de l'enquêteur, le budget limité retenu pour le remboursement des déplacements, le nombre de visites supplémentaires à prévoir compte tenu de la saison estivale et le délai de réalisation du projet. Nous avons dû fixer la taille de l'échantillon à 225 ménages. Ainsi la probabilité de sélection d'un ménage était de 4,4%.

Afin de se convaincre que cette taille d'échantillon est raisonnable, calculons l'erreur relative maximale que nous imputera cette taille d'échantillon. Avec une taille d'échantillon de 225 logements, le théorème suivant montre que l'erreur relative pour une proportion autour de 0,5 est de 12,78%. Le théorème montre également que les tailles minimales d'échantillons requises pour obtenir des erreurs relatives de l'ordre de 5% ou de 10%, sont respectivement de 1183 et de 358 logements. Ainsi, pour obtenir un gain de précision de 2,78%, il aurait fallu que l'enquêteur visite 133 ménages supplémentaires, ce qui était totalement impossible compte tenu du temps et du budget.

THÉORÈME 4.1.1

L'erreur absolue requise pour avoir un intervalle de confiance de niveau 1-α pour l'estimation d'une proportion est approximativement égale à :

$$\varepsilon = \frac{z_{\alpha/2}\sqrt{N-n}}{2\sqrt{n(N-1)}}.$$

Démonstration :

On doit avoir

$$z_{\alpha/2}\sqrt{\frac{P_y(1-P_y)(N-n)}{n(N-1)}} \leq \varepsilon.$$

Or, cette borne inférieure de ε est maximale lorsque la fonction $P_y(1-P_y)$ est maximale. Le maximum de cette fonction est atteint en $P_y = 0,5$.

D'où

$$z_{\alpha/2}\sqrt{\frac{0,5(1-0,5)(N-n)}{n(N-1)}} = \frac{z_{\alpha/2}\sqrt{N-n}}{2\sqrt{n(N-1)}} = \varepsilon$$

■

Pour notre taille d'échantillon de 225 unités issue d'une population de 5126 unités, l'erreur absolue associée à un intervalle de confiance pour une proportion est de 6.39% et ce, 19 fois sur 20. Il nous reste finalement à déterminer la répartition des 225 ménages au travers des trois strates. Mais, comme nous avons déjà choisi l'affectation de type identique, la taille d'échantillon qui sera prise dans chacune des strates sera de 75 ménages.

4.2 L'algorithme de tirage

Cette section présente une dissection de l'algorithme de tirage qui a permis la sélection de l'échantillon dans la base de sondage. Plusieurs algorithmes concurrents existent. Parmi les algorithmes les plus populaires, on retrouve le tirage à l'aide d'une table de nombres pseudo-aléatoires et le tirage systématique. On résume la première méthode en soulignant qu'elle consiste dans un premier temps à numéroter les individus et ensuite à utiliser une table de nombres pseudo-aléatoires pour générer n nombres au hasard différents et par conséquent, sélectionner les n individus qui constitueront l'échantillon. La seconde méthode, plus rapide, consiste à ranger les N individus de la population selon un certain ordre; pour obtenir un échantillon de taille n, on choisit au hasard un nombre k parmi les K premiers, et ensuite on choisit les individus subséquents à toutes les K unités. Pour cette enquête, nous avons choisi d'utiliser la première méthode, car elle correspond au schéma du plan aléatoire simple ou stratifié.

Pour schématiser cette méthode de tirage, nous avons besoin de trois étapes:

L'ALGORITHME DE TIRAGE

Étape 1 :

Avant d'effectuer le tirage, il faut créer les trois strates. Pour ce faire, nous avons dans un premier temps divisé notre population totale selon les critères de définition exclusive des strates, en l'occurrence, le nom des rues. Après une visite

sur le terrain et à l'aide d'une carte géographique, nous avons décidé quelle rue était incluse dans chacune des trois strates.

Étape 2 :

Pour une strate donnée, nous avons numéroté chacun des bâtiments habitables de 0 à $N_h - 1$.

Étape 3 :

Nous avons tiré notre échantillon à l'aide de nombres pseudo-aléatoires extraits d'une table génératrice de nombres aléatoires. Pour l'extraction, nous avons suivi la procédure qui suit, elle se divise en cinq opérations récursives.

0. (Initialisation de l'algorithme) Sans regarder, déposez au hasard un doigt sur la table des nombres et choisir le nombre immédiatement sous celui-ci. Lire les trois premiers chiffres de celui-ci et passez à l'opération 2.
1. Lire en ligne dans la table les trois chiffres qui suivent immédiatement et ce, malgré les espaces entre les valeurs de la table. Si vous êtes rendu au bout de la ligne, poursuivre la lecture au début de la ligne suivante. Si vous êtes rendu à la fin de la table, poursuivez votre lecture au début de la première ligne du haut de la table. Passez à l'opération 2.
2. Divisez ce nombre de trois chiffres par la taille de la population N_h de la strate où on effectue le tirage. Passez à l'opération 3.
3. Isolez le reste, r, de la division. Prenez note que ce reste, r, sera nécessairement compris entre 0 et $N_h - 1$. Passez à l'opération 4.
4. Sélectionnez l'individu dont le numéro est égal au reste. Si l'individu s'avère avoir été sélectionné, répétez depuis l'opération 1 aussi longtemps qu'un reste corresponde à un individu non sélectionné. Sinon, Passez à l'opération 5.
5. Répétez les opérations 1 à 5 jusqu'à ce que le nombre d'individus sélectionnés corresponde à la taille de l'échantillon désirée.

Cette procédure est efficace pour toute table de nombres pseudo-aléatoires. Celle qui a été utilisée pour cette enquête est celle incluse à la page 622 de [4]. De plus cet algorithme

permet d'effectuer un tirage probabiliste, c'est-à-dire un tirage où chaque individu de la population possède une probabilité connue à l'avance, n_h / N_h, d'être sélectionné.

Chapitre 5

La conception du questionnaire

Ce chapitre se veut être une suite logique au premier chapitre dans lequel nous avons déterminé les objectifs de l'enquête. Dans les préparatifs d'une enquête, la conception du questionnaire est élaborée en parallèle avec les étapes de la recherche de la base, de la sélection du plan d'échantillonnage et du tirage de l'échantillon.

Dans ce chapitre nous présentons les différentes étapes de l'élaboration d'un questionnaire qui se trouve être la dernière étape qui nous sépare de la collecte des données. Cette dernière est cruciale dans le sens où la précision et la validité des résultats en dépendent directement. Dans les sections qui suivent, nous y avons adapté le contenu du guide docimologique du ministère de l'éducation du Québec intitulé : "Conseils pratiques pour la construction d'un instrument de mesure" [5]. Ce fascicule, le cinquième de la série formant le guide docimologique, a pour but d'aider les agents d'éducation à construire des instruments de mesure.

Pour discuter de toutes les facettes entourant l'élaboration d'un questionnaire, nous avons besoin de cinq sections. Dans un premier temps, inspiré du fascicule du MEQ [5], nous présenterons une classification des différents types de questions qu'il est possible de retrouver dans un questionnaire. Dans un second temps, nous nous attarderons sur quelques conseils généraux quant à la formulation des questions. Dans un troisième temps, nous présenterons les différentes étapes consacrées à l'assemblage du questionnaire. Dans un quatrième temps, nous présenterons l'étape de la pré-enquête comme faisant partie intégrale de la conception du questionnaire. Dans un cinquième et dernier temps, nous présenterons et critiquerons les questions utilisées dans notre enquête.

5.1 Les types de questions

La préparation des questions n'est pas une tâche facile. Mais les efforts qu'elle exige seront largement compensés par la qualité des résultats obtenus. Il existe plusieurs façons de classer

les questions. La pratique courante est de les classer en fonction de l'interprétation que le responsable de l'enquête aura à faire des réponses que le répondant aura fournies. Plus précisément, il y a deux grandes catégories de questions : les questions dont les réponses doivent être interprétées de façon objective et celles dont les réponses doivent être interprétées de façon subjective.

Une question à interprétation objective, que l'on associe trop souvent à une question à choix multiples, est caractérisée par le fait qu'elle est limitée et porte sur un aspect bien précis. En général, l'ensemble des réponses possibles est connu et déterminé à l'avance par le rédacteur. Par exemple, le répondant fournit une réponse courte ou choisit une des réponses suggérées. La plus grande faiblesse des questions à interprétation objective réside dans le fait que l'enquêté peut deviner la réponse ou répondre au hasard. À l'opposé, une question à interprétation subjective, que l'on associe généralement à une question à développement ou encore ouverte, est caractérisée par un énoncé qui est plus vaste que précis. Les réponses sont en partie déterminées par le rédacteur mais il ne peut prévoir la façon dont les répondants le formuleront. L'enquêté organise sa réponse et l'exprime dans ses propres mots. Cependant, le sujet peut dire n'importe quoi à l'enquêteur.

Les questions à interprétation objective sont efficaces pour mesurer la connaissance d'un fait. Elles sont moins efficaces pour mesurer des habiletés comme exprimer des idées ou résoudre des problèmes. Les statisticiens disposent de bon nombre d'outils de travail pour décrire et explorer les réponses à ce type de question. Les questions à interprétation subjective, sont particulièrement efficaces pour mesurer des comportements marginaux. Mais ces dernières sont moins efficaces pour mesurer des faits particuliers. En dehors de l'analyse lexicale, les statisticiens disposent de bien peu d'outils pour interpréter ce type de réponses qui se présentent en général sous forme de texte écrit à la main.

5.1.1 Les questions à interprétation objective

En nous inspirant du fascicule 5 du guide docimologique, nous traitons et adaptons ici les différents types de questions à interprétation objective qui peuvent être utilisées lors de l'élaboration d'un questionnaire. Parmi les plus courantes, on retrouve la question à réponse courte, la question à choix multiples, le regroupement de connaissances, l'exercice d'appariement et finalement l'alternative. Dans chacun des cas, nous allons proposer une

définition, illustrer les principales formes, donner des indications sur son utilité et suggérer des principes de rédaction.

5.1.1.1 La question à réponse courte

Il s'agit de questions dont le répondant répond par un mot, un groupe de mots ou un nombre. Dans ce type de question il y a toujours un espace réservé pour inscrire la réponse. Cette dernière tâche peut être effectuée soit par l'enquêteur, soit par l'enquêté. Ce type de question peut prendre différentes formes. Par exemple on peut poser une question directe, demander de compléter une phrase ou demander de trouver un élément dans un tableau et d'inscrire la réponse dans l'espace réservé à cet effet.

Ce type de question est surtout utile pour obtenir de l'information sur un fait particulier. Elle fait surtout appel à la mémoire et nécessite que l'enquêté soit capable de trouver une réponse par lui-même. L'avantage relié à ce type de question tient du fait qu'elle exige une réponse du répondant ce qui limite avantageusement les réponses dues au hasard. Ses limites sont dues au fait qu'elles sont difficiles à construire. En effet, d'une part, il faut éviter toute ambiguïté quant à la nature de la réponse attendue et, d'autre part, il faut s'assurer qu'il n'y a qu'une seule bonne réponse possible ou bien, être en mesure de les prévoir toutes.

Exemples :
- Question directe :

  Quel est votre âge?: _____
- Question directe qui demande d'effectuer une opération avant de fournir la réponse :

  Additonnez les lignes 235 et 245: Total : _____

Pour les questions à réponses courtes, certains principes de rédaction doivent être suivis. Premièrement, il ne faut recourir à ce type de questions que dans les cas où il est possible de répondre par un mot, une expression, un nombre ou un symbole. Deuxièmement, s'assurer que l'énoncé est exempt de toute ambiguïté. L'enquêté doit savoir exactement sur quoi porte la question. Troisièmement, il faut s'assurer qu'il n'y a qu'une seule réponse possible pour chacun des individus. Quatrièmement, il est préférable de s'arranger pour que les réponses soient exprimées en chiffre rond. Cinquièmement, la pratique a montré qu'il vaut mieux

utiliser des questions directes plutôt que des phrases à compléter. Finalement, si un énoncé exprime une opinion plutôt qu'un fait, il faut attribuer cette opinion à quelqu'un.

5.1.1.2 La question à choix multiples

La question à choix multiples comporte une partie initiale (le tronc) qui peut prendre la forme d'une question directe ou d'un énoncé incomplet, et un certain nombre de réponses suggérées. On retrouve ce type de questions sous plusieurs formes. Soit une seule réponse, plusieurs réponses possibles, le classement selon un ordre de préférence donné, les échelles d'appréciation et la substitution.

De tous les types de questions à interprétation objective, la question à choix multiples est celle qui se prête au plus grand nombre d'applications. Elle permet de mesurer la connaissance d'un fait particulier, d'évaluer le degré de satisfaction, de classer certains produits selon un ordre de préférence, etc. À son avantage, elle offre une flexibilité remarquable tout en offrant une grande objectivité. Par ses choix de réponses, elle permet une plus grande caractérisation des phénomènes que l'on tente de mesurer. Cependant ces questions sont difficiles à construire. La difficulté réside d'une part dans la formulation de la question et, d'autre part dans l'identification de toutes les modalités pertinentes reliées au phénomène. Ce dernier problème peut être résolu à l'aide de la pré-enquête, il suffit de poser la question sous la forme de réponse courte et l'ensemble des réponses formaliseront les diverses modalités qui seront utilisées lors de l'enquête.

Exemples :

- Plusieurs bonnes réponses :

Quels sont (ou seraient) vos critères de sélection d'une firme d'expert-conseils? Parmi les 15 critères suivants, choisir et cocher les 4 critères les plus importants selon vous :

( ) Ponctualité	( ) Jeunesse	( ) Rapidité
( ) Expérience	( ) Qualité du travail	( ) Compétence
( ) Professionnalisme	( ) Crédibilité	( ) Nouveauté
( ) Coût	( ) Respect des délais	( ) Formation
( ) Service personnalisé	( ) Accessibilité	( ) Autres : _____

- Échelle d'appréciation de type graphique :

Placez un "x" sur la partie de l'échelle qui, selon vous, représente le mieux votre degré (ou niveau) de satisfaction face au transport en commun de votre région.

+----------+----------+----------+----------+

très faible faible moyen bon excellent

Pour construire ce type de question, il faut suivre quelques principes de rédaction. D'une part, on doit tenir compte de l'énoncé. Ainsi, si l'on pose une question directe, il faut s'assurer que l'énoncé présente un fait précis et qu'il est complet en lui-même. On doit pouvoir comprendre le sens de la question sans être obligé de lire les choix de réponses. Si l'on a recours à une phrase à compléter, il faut vérifier qu'il n'y a pas d'inconsistance grammaticale entre l'énoncé et chacune des réponses suggérées.

D'autre part, on doit tenir compte des réponses suggérées. Toutes les réponses doivent être de même nature. Aussi, toutes les réponses doivent être sensiblement de la même longueur. Il faut identifier les choix de réponses avec des lettres au lieu de chiffres. Cette stratégie évite la confusion avec les numéros des questions. Si possible, essayer de disposer les réponses suggérées en colonnes, de façon à former un bloc dégagé du tronc de l'énoncé. Finalement, ne pas abuser des formules du type : "aucune idée", qui motivent les gens à ne pas répondre.

5.1.1.3 Le regroupement de connaissances

Ce type de question à interprétation objective exige du répondant à replacer dans un ordre donné (selon l'ordre logique, selon l'ordre chronologique, par le goût ou simplement par ordre de préférence) une série d'énoncés ou d'éléments présentés dans un désordre quelconque. Cette ordre peut être indiqué par des lettres ou des chiffres. Ce type de question peut se présenter sous deux formes différentes, la forme de base où le répondant réarrange lui-même l'ordre des éléments et la forme limitée par un choix de réponses fixes.

L'utilité d'une question à réarrangement réside dans le fait qu'elle est capable, avantageusement, de faire ressortir les préférences chez les individus. Elle est très utile en marketing. Cependant, l'interprétation des résultats peut être fastidieuse ou peu

représentative si les gens sont indécis ou si les gens ne connaissent pas un ou plusieurs des produits en question.

Exemple :

Classez les marques de bières suivantes selon votre ordre de préférence :

1. Molson Dry	4. La Maudite
2. Labatt 50	5. La Tremblay
3. Streingnen	6. Black Label

Réponse : _____ _____ _____ _____ _____ _____

préférée --> moins préférée

Pour rédiger ce type de question, il ne faut placer que des énoncés qui sont du même type. De plus, il est impératif de déterminer l'ordre au hasard. Évitez les modalités inconnues qui n'apportent que de la confusion.

5.1.1.4 L'appariement

L'exercice que propose ce type de question prend ordinairement la forme de deux listes d'éléments qui doivent être associés entre eux selon une règle donnée. Il s'agit possiblement de la forme que l'on rencontre le moins dans un questionnaire de sondage. Elle présente plus d'inconvénients que d'avantages. Ces questions sont lourdes et longues, ce qui va à l'encontre d'un questionnaire efficace pour un sondage. Elle s'adresse plus facilement à des échantillons spécifiques d'individus comme en témoigne le premier exemple :

Exemple :

Dites quel sera l'effet de chacune des mesures économiques énumérées ci-dessous. Inscrivez à côté de chaque mesure la lettre qui correspond à l'effet que vous jugez le plus probable.

A- augmentera les inégalités de revenus

B- diminuera les inégalités de revenus

C- n'affectera pas les inégalités de revenus

1. Augmentation progressive de l'impôt sur le revenu _____
2. Introduction d'un impôt spécial sur les gains en capitaux _____
3. Introduction d'une taxe de vente sur les vêtements _____

Exemple :

Quels sont ou seront (seraient) les trajets-type des membres de votre famille? Reliez à l'aide d'un trait un départ à une destination.

dép.	dest.
dép.	dest.
dép.	dest.
dép.	dest.

5.1.1.5 L'alternative

Ce dernier type de question comporte deux réponses possibles entre lesquelles le répondant doit choisir (vrai/faux, oui/non, fait/opinion, d'accord/pas d'accord, etc.). Cette forme de question remplace avantageusement les questions à choix multiples lorsqu'il n'y a que deux réponses possibles. L'alternative est possiblement la forme de question la plus puissante que l'on peut retrouver dans un questionnaire d'enquête.

L'alternative trouve toute son utilité dans les questionnaires de sondage d'opinion car elle nécessite une réponse courte. Elle se prête avantageusement au calcul des proportions dans une population. Ce caractère dichotomique lui confère d'ailleurs un statut très particulier. Un seul inconvénient survient, il faut être en mesure d'élaborer la question de façon à ce que deux seules réponses soient possibles.

Exemple :

Le modèle classique :

Votre entreprise a-t-elle, actuellement, recours à des services externes d'expert-conseils en ingénierie mécanique OU en statistiques appliquées?

( ) Oui

( ) Non

Les principes de rédaction sont simples. Il faut utiliser des énoncés qui n'ont que deux seules réponses possibles en toutes circonstances. Il faut éviter toutes négations.

5.1.2 Les questions à interprétation subjective

Selon le guide docimologique [5], on distingue deux types de question à interprétation subjective. Selon la longueur et la complexité de la réponse attendue, on parle de questions à réponses limitées ou de questions à réponses élaborées.

5.1.2.1 La question à réponse limitée

Ce sont les questions dont le champ est assez restreint et qui appellent une réponse relativement courte; une ou deux phrases généralement. On utilise en général ce type de questions dans les enquêtes préliminaires. Elles permettent de déterminer l'ensemble des réponses possibles qui serviront éventuellement à construire les questions à choix multiples. Elles sont simples à construire et il suffit de restreindre à deux ou trois lignes l'espace pour la réponse.

Exemple:

Pour des raisons d'agrandissement, la ville de Rock Forest est dans l'obligation de déménager la bibliothèque, avez-vous des suggestions quant à l'endroit où il serait préférable de la déménager.

_____

_____

_____

_____

_____

5.1.2.2 La question à réponse élaborée

Ce type de question porte sur des problèmes plus complexes qui exigent des efforts de réflexion et d'organisation. Ce type de question est plutôt rare dans les sondages d'opinion conventionnels. On les retrouve dans les entrevues au travers desquelles on isole le répondant qui s'exprime de façon libre sur les différentes questions du sondeur. Les lettres d'opinions libres en sont un exemple.

Le principal avantage relié à ce type de question est qu'il permet l'élaboration ainsi que la confrontation des idées. Le problème que soulève ce type de question réside dans la difficulté de l'évaluation objective de l'opinion. À ce titre la formation du sondeur est très importante.

Exemple:

Que pensez-vous de l'avortement?

_____

...

_____

Ne recourir à ce type de question que pour les comportements qui ne peuvent être mesurés adéquatement par un autre type de question. Formuler la question de façon claire et précise, le répondant doit savoir exactement ce que l'on attend de lui.

5.2 La formulation des questions

Avant de commencer à rédiger les questions, il faut se demander d'une part quel est le type d'information que l'on cherche, quel attribut ou quelle particularité on veut mesurer. Et plus particulièrement sous quelles formes de comportement ou d'habitude cet attribut ou cette particularité pourrait se manifester. D'autre part, quel type d'activité ou de question permettrait de déceler si cet attribut ou cette particularité est présente ou absente chez un individu.

Le guide docimologique [5] souligne l'importance de ne poursuivre qu'un seul but par question. Il précise aussi qu'il faut rédiger les questions à l'avance et en rédiger plus qu'il n'en faut afin de pouvoir opérer un choix au moment de l'assemblage. Il faut éviter tous les stéréotypes de type culturel, ethnique, religieux, social ou sexiste qui sont des sources de biais systématiques.

Au-delà des distorsions précédentes, le guide rappelle que certains mots, ou groupement de mots sont à éviter, par exemple il faut fuir les négations, les doubles négations ou encore les mots relatifs comme toujours, trop, seulement, jamais, aucun, tous, parfois, en général,

ordinairement, etc. Il faut aussi surveiller toutes les inconsistances grammaticales, sémantiques, orthographiques et typographiques.

Il est recommandé de formuler les questions en fonction de la population à qui elles s'adresseront. Il faut aussi que les directives soient suffisamment claires et précises pour que toute personne qui possède les connaissances ou les habiletés à mesurer puisse répondre correctement. Finalement, il est recommandé de faire relire les questions par des collègues afin d'en vérifier la pertinence et la qualité technique de chacune.

5.3 L'assemblage du questionnaire

L'assemblage du questionnaire est une autre étape importante de l'élaboration du questionnaire. En effet, la somme des durées moyennes des réponses à chacune des questions définit le temps moyen de réponse au questionnaire qui ne doit pas être trop long. L'agencement des questions doit être tel qu'il ne provoque pas chez le répondant un effet négatif qui modifierait ses réponses aux questions subséquentes.

5.3.1 Le choix des questions

Parmi l'ensemble des questions que l'on a préparé, on opère un choix en tenant compte de certains critères. Tout d'abord, il faut toujours avoir en tête que l'ensemble de l'épreuve doit correspondre à l'ensemble des comportements que l'on veut évaluer. Ensuite, il faut tenir compte de la durée moyenne de remplissage du questionnaire. Il faut bien comprendre que différents facteurs tels la fatigue, la baisse de l'intérêt ou de concentration, la hâte de finir, parmi bien d'autres, risquent d'invalider les résultats d'un questionnaire trop long. Finalement, le niveau du vocabulaire ainsi que le degré de difficulté pour l'ensemble du questionnaire se doivent d'être adaptés à la population à laquelle il s'adresse.

5.3.2 L'agencement des questions

Le problème de l'agencement des questions se pose surtout pour les épreuves qui comportent un assez grand nombre de questions. Il faut tenir compte des différents types de questions, du niveau de complexité des questions prises une à une et prises toutes ensemble et, enfin, de la nature des comportements mesurés.

Le guide docimologique [5] propose une démarche que nous présentons maintenant. Dans un premier temps, il convient de regrouper les questions de même type. Ceci a pour avantage de minimiser le nombre de consignes à donner au répondant. Dans un second temps, on agencera les questions de façon à les présenter en allant des plus simples aux plus complexes. Dans un dernier temps, dans chacune des sections, on regroupera les questions selon le type de comportement à mesurer.

Pour ce qui est de la présentation du questionnaire, il est impératif qu'elle soit aérée et uniformisée. Il faut éviter les recto-verso, ce qui permet de minimiser le nombre de non-réponse par inadvertance du répondant.

5.3.3 La révision du questionnaire

Il est recommandé de soumettre le questionnaire à une personne qui n'a pas participé à l'élaboration de l'outil pour une vérification de contenu. Cette étape est nécessaire puisqu'elle permettra de déceler les dernières ambiguïtés qui auraient échappé au rédacteur.

Nous proposons deux grilles de révision. La première grille effectue la révision de chacune des questions qui forme le questionnaire. La seconde grille a été élaborée pour effectuer la révision du questionnaire dans sa globalité.

Tableau 1 : Grille de révision des questions

Pour chaque question	oui	non	incertain	ne s'applique pas
Exige-t-elle des connaisances bien définies?				
Mesure-t-elle un caractère bien précis?				
Demande-t-elle un fait précis?				
Correspond-t-elle à l'objectif qu'elle doit mesurer?				
Les connaissances ou les attributs qu'elle mesure sont-ils suffisamment importants pour qu'on les mesure?				
Le vocabulaire est-il adapté à la population à laquelle il s'adresse?				
La syntaxe est-elle adaptée au niveau de développement de la population à laquelle elle s'adresse?				

A-t-on éliminé dans l'énoncé les renseignements superflus?				
Les consignes sont-elles suffisamment claires pour qu'un individu qui possède les connaissances ou l'attribut puisse répondre correctement?				
Cette question est-elle indépendante des autres?				
Est-elle exempte de toute distorsion susceptible de biaiser les résultats en favorisant (ou défavorisant) certains individus plus que d'autres: absence de biais religieux? biais culturel? biais ethnique? biais social? biais sexiste?				
Est-elle exempte de tout stéréotype... culturel? ethnique? social? sexiste?				
La source des textes ou opinions est-elle indiquée?				
Le temps de remplissage de cette question est-il adéquat?				
Les principes de rédaction propres à ce type de questions sont-ils respectés?				
La réponse à la question est-elle dépendante de la saison courante?				

Tableau 2 : Grille de révision du questionnaire

Pour l'ensemble du questionnaire	oui	non	incertain	ne s'applique pas
Le but de l'enquête est-il clairement établi?				
Y a-t-il concordance entre le contenu de l'épreuve et les objectifs que l'on veut évaluer?				
Le choix du (ou des) type(s) de question(s) est-il judicieux?				
L'agencement des questions est-il pertinent?				
La longueur de l'épreuve est-elle appropriée?				

Les directives sont-elles... claires? complètes? précises? bien en évidence?				
Les consignes pour répondre aux questions sont-elles... claires? complètes? précises? bien en évidence?				
Le niveau de langage (vocabulaire et syntaxe) est-il adapté à la population à qui le questionnaire s'adresse?				
Le questionnaire est-il exempt d'erreurs... grammaticales? orthographiques? typographiques?				
Les caractères d'imprimerie sont-ils clairs?				
L'ensemble du questionnaire est-il uniformisé?				

Le questionnaire qui a été utilisé pour l'enquête est fourni en annexe A.

5.4 La pré-enquête

La pré-enquête, souvent mise de côté, est la phase la plus importante du processus d'élaboration du questionnaire. Cette phase permet de valider le questionnaire. Plus précisément, cette phase permet de calibrer l'interaction entre notre questionnaire et la population cible. Des ajustements pourront être apportés ensuite.

Lors de la validation, différents types de situations peuvent se présenter. Pour chacune de celles-ci, il est possible de faire des ajustements appropriés. Nous résumons ces diverses situations ainsi que les ajustements associés dans le tableau suivant.

Tableau 3 : Grille de validation d'un questionnaire

Situation possible	Ajustement à faire
Plus de cinq fois, le sondeur a été obligé d'expliquer la question avant qu'elle ne soit comprise.	- Revoir la formulation de la question - Corriger la question ou en élaborer une autre
Tout le monde a fourni la même réponse à une question.	- Rejeter la question - Revoir à quel paramètre est reliée cette question - La question était-elle pertinente? - La réponse était-elle suggérée? - Élaborer une nouvelle question sous ces considérations
Le sondeur ne comprend pas la réponse donnée.	- Le sondeur note la réponse - On ajoute une modalité s'il s'agit d'un choix de réponses - On ajoute une nouvelle question couvrant cette situation
Une question met en doute la crédibilité de l'étude.	- Retirer la question.
La réponse à une question possède une influence marquée sur les réponses subséquentes.	-Revoir la formulation de la question. -Revoir l'ordre des questions influençables. -Modifier l'ordre ou changer la question.
On discerne une faute d'orthographe.	- On corrige la faute.
Certaines personnes sont offusquées par une question.	- Revoir la question, nous sommes possiblement en présence d'un biais de type religieux, ethnique, etc.
Les gens sont exaspérés.	- Revoir la durée du questionnaire.
Une question est non pertinente.	- Modifier la question ou en écrire une autre.

Il est important que le responsable du projet participe à cette phase. Cette expérience lui permettra de mieux rédiger les consignes aux futurs sondeurs, ce qui limitera les erreurs d'observation dues au sondeur lui-même. Pour notre pré-enquête, nous avons sélectionné une soixantaine de familles. Cette étape nous a permis d'ajuster, d'ajouter et de supprimer certaines questions. Une dernière mise en garde doit être signalée, si trop de corrections sont apportées au questionnaire, une seconde étude préliminaire est fortement recommandée. Celle-ci permettra de mesurer l'interaction entre les questions initiales et les corrections ajoutées. Nous discuterons de toutes ces facettes dans la prochaine section.

5.5 Les questions de l'enquête

Dans cette section nous discutons de la conception du questionnaire qui a servi à notre enquête. Nous aborderons plusieurs des thèmes déjà discutés dans les sections précédentes afin d'en montrer le côté pratique. Nous prendrons aussi conscience qu'aucun détail ne peut être laissé au hasard et que l'on ne peut garder sous aucun prétexte une question jugée seulement acceptable. En somme un questionnaire se doit d'être pratiquement parfait, du moins en théorie.

Le questionnaire final a été présenté à 225 familles et comportait deux parties bien distinctes. Ainsi, nous allons tout d'abord présenter la première partie du questionnaire qui correspondait à l'étude sur le transport en commun. Ensuite, nous allons faire de même avec la deuxième partie du questionnaire qui correspondait à l'étude sur l'avenir de la bibliothèque municipale.

5.5.1 Le questionnaire sur le transport en commun

L'étude sur le transport en commun avait pour but ultime de dégager la satisfaction des usagers face au transport collectif. Pour atteindre cet objectif, neuf questions ont été nécessaires. Nous allons jeter un regard attentif sur chacune de ces questions.

Plus précisément, pour chacune d'entre elles, nous allons tenter, si possible, de répondre aux questions suivantes : <<Quelle est l'objectif que doit poursuivre la question courante?>>, <<Quel est le type de question qui permettrait d'obtenir l'information voulue?>>, <<La formulation de la question satisfait-elle les principes de rédaction?>>, <<Quelle a été l'interaction entre les gens et la question lors de la pré-enquête ou lors de l'enquête?>> et finalement, <<A posteriori, quels seraient les correctifs à apporter s'il y en a?>>.

5.5.1.1 La question #1

L'objectif premier relatif à la première question de tout questionnaire est de pouvoir mettre en évidence les différents états possibles d'un ménage interrogé. Initialement, nous avons compté deux états possibles pour un ménage : celui-ci utilise le transport en commun ou ne l'utilise pas. Pour les fins de l'étude, nous avons établi les deux définitions suivantes : les familles clientes sont les familles qui utilisent, de près ou de loin, l'autobus comme moyen de

transport et les familles indépendantes sont celles qui ne ressentent pas le besoin d'utiliser ce mode de transport.

Pour dégager une proportion, une question à alternative est l'outil idéal. De plus, ce type de question revêt d'autres avantages tels la rapidité, la clarté et la facilité des calculs des estimateurs. Pour toutes ces raisons, nous avons choisi une question alternative pour débuter le questionnaire.

Q1. Est-ce que vous ou quelqu'un de votre famille prenez, ou pensez prendre l'autobus comme moyen de transport?

- oui
- non

La formulation de la présente question respectait a priori tous les principes de rédaction relatifs aux questions de ce type. Cependant, l'étude préliminaire a montré la présence d'un troisième état possible pour un ménage. En effet, une famille pouvait être cliente actuellement sur le réseau, indépendante du réseau de transport ou cliente potentielle. Suite à l'étude préliminaire, nous avons dû définir les familles potentielles comme étant celles qui n'utilisent pas et n'utiliseront pas le service de transport en commun sous sa forme actuelle, mais qui, si quelques modifications sont apportées au réseau, deviendraient possiblement des familles clientes.

Pour résoudre cette problématique, deux solutions étaient possibles. Soit faire de la première question une question à choix multiples en ajoutant la modalité "potentielle", soit ajouter une nouvelle question à alternative couvrant l'état des familles potentielles. Pour des raisons de clarté et de rapidité, nous avons opté pour la deuxième solution. La question 2 qui suit a ainsi été ajoutée.

Par la suite, l'interaction entre l'ensemble des répondants et la question a été excellente. Autant les jeunes personnes que les plus âgées ont pu comprendre le sens de la question. Aucune ambiguïté n'a été soulevée concernant l'état à laquelle appartenait un ménage. A posteriori, ne pas avoir modifié cette question au profit d'une à choix multiples semble avoir été un bon choix.

5.5.1.2 La question #2

L'objectif à atteindre pour cette seconde question est de dégager la proportion de familles potentielles dans la population. Comme il s'agit de mettre en évidence une proportion et que nous avons fait le choix de ne pas transformer la première question en choix multiples, une question alternative s'impose.

Q2. Si non, supposons que l'autobus soit plus accessible, est-ce que vous ou quelqu'un de votre famille prendriez l'autobus comme moyen de transport?

- oui
- non

L'ajout de cette question a finalement permis de mettre en évidence les proportions des trois états possibles de la population: famille cliente (Q1=oui, Q2= nil), famille potentielle (Q1=non, Q2= oui) et finalement famille indépendante (Q1=non, Q2= non). La formulation de cette question respecte tous les principes de rédaction relatifs à ce type de question. De plus, elle respecte le principe d'agencement des questions qui consiste à regrouper les questions semblables.

L'interaction entre les répondants et cette question a été excellente. Cette question a apporté la solution au problème déjà soulevé par la question précédente. Notons qu'un ménage qui répondait "non" aux deux premières questions mettait fin instantanément à cette partie du questionnaire. Il est arrivé à quelques reprises que le sondeur se soit trouvé en présence de personnes qui ne voulaient pas répondre. La rapidité et la forme des questions ont permis au sondeur d'obtenir tout de même les réponses aux deux premières questions qui étaient très souvent non et non. En d'autres mots, la rapidité et le caractère dichotomique des questions à alternatives, ont sûrement limité le taux de non-réponses qu'un autre type de question aurait provoquées. Ce qui nous amène à dire qu'il est très important qu'un questionnaire puisse cibler dans quel état un individu se trouve dès les premières questions. Si cet individu ne veut pas répondre, nous savons au moins quel état le caractérise.

5.5.1.3 La question #3

Une fois qu'une famille est identifiée comme étant cliente ou potentielle, l'objectif de la troisième question est d'obtenir une estimation du total des individus de la population (d'individu unique et non les ménages) qui prennent ou pensent prendre l'autobus comme moyen de transport. Pour être capable de tirer un estimateur, il faut connaître le nombre de personnes dans chacune des familles qui prennent ou pensent à prendre l'autobus. Pour couvrir l'ensemble des situations possibles, nous avons proposé une question à choix multiples.

Q3. Combien de personnes de votre famille prennent ou pensent prendre (prendraient) l'autobus?

- 1
- 2
- 3
- 4 et plus

A priori la formulation de cette question respectait les principes de rédaction relatifs à ce type de question. L'étude préliminaire n'a démontré aucune anomalie d'interaction. Pourtant, l'analyse des résultats a démontré que cette question n'avait possiblement pas le même sens pour les familles clientes que pour les familles potentielles. En effet, nous avons pu constater que la majorité des familles clientes n'ont qu'un seul membre actif sur le réseau de transport en commun, alors que ce n'est pas le cas pour les familles potentielles où la majorité a fait état que deux de leurs membres seraient sujets à prendre l'autobus comme moyen de transport.

L'origine de cette situation s'explique par le fait que l'état des familles potentielles n'a été découvert qu'après la pré-enquête et qu'une seconde étude préliminaire n'a pas été faite afin de tester l'interaction entre les questions de départ et ce nouvel état de la population. A posteriori une seconde étude préliminaire aurait pu éviter ce genre de problème.

5.5.1.4 La question #4

L'objectif associé à cette question consiste à connaître les régularités en terme de fréquences de passage pour l'ensemble des familles. Trois modalités ont été recensées: régulier, semi-régulier ou occasionnel. Connaissant l'univers des réponses, une question à choix multiples s'avère l'outil idéal dans cette situation.

Q4. Le membre de votre famille qui utilise, ou utilisera (utiliserait) le plus souvent le réseau de transport en commun est ou sera (serait) un passager de type

- régulier (plus de 10 passages par semaine)
- semi-régulier (entre 6 et 9 passages par semaine)
- occasionnel (entre 0 et 5 passages par semaine)

La formulation de cette question satisfait les principes de rédaction de base. De plus, les précisions apportées entre parenthèses éclairent sur le sens des mots "réguliers" et "occasionnels" en les associant à des valeurs quantitatives. L'interaction entre cette question et la population a été excellente compte tenu du fait que les termes régulier, semi-régulier et occasionnel étaient des termes qui illustraient très bien les différents comportements étudiés ici. De fait, aucune fois le sondeur n'a eu à préciser davantage les termes de la question.

L'analyse des résultats nous a permis de constater que les gens se définissent peu comme étant des usagers de type semi-réguliers. En fait, par leur fréquence de passage (entre 6 et 9 par semaine), les semi-réguliers ont plus de chance de se comporter comme des réguliers (achat d'une passe par exemple). Ainsi dans les calculs des estimateurs, les semi-réguliers ont été ajoutés aux réguliers. A posteriori, cette question resterait la même car elle permet d'isoler les semi-réguliers généralement des familles qui ne savent pas si elles devraient ou non acheter une passe.

5.5.1.5 La question #5

L'objectif que poursuit cette question consiste à faire ressortir les points chauds du réseau. Pour cela, il faut connaître où vont les gens. Comme il est impossible d'établir une liste de toutes les destinations, une question à réponse courte disposant de plusieurs espaces de réponse est le meilleur outil disponible pour recueillir ce type d'information.

Q5. Quels sont ou seront (seraient) les trajets-type des membres de votre famille?

dép. dest.

dép. dest.

La formulation de cette question répond aux divers critères de rédaction. La diversité des réponses à cette question a permis la création d'un répertoire de destinations. Pour les familles clientes, il s'agit des destinations où elles se rendent présentement. Tandis que pour les familles potentielles, il s'agit plutôt des destinations où elles aimeraient aller.

L'interaction entre cette question et la population a été relativement bonne. D'une part, elle a été un peu ambigüe pour les personnes qui répondaient au nom d'une autre personne de sa famille absente au moment de l'enquête; d'autre part, certaines réponses n'ont pu être traitées compte tenu du fait qu'elles étaient trop particulières ou encore trop vagues. Par exemple, des gens ont répondu qu'ils se rendaient à l'intersection des rues Cabana et Galt tandis que d'autres se rendaient simplement à Sherbrooke. A posteriori, connaissant l'ensemble des destinations les plus populaires de cette population, cette question pourrait éventuellement devenir une question de type appariement dans une étude similaire sur la même population.

5.5.1.6 La question #6

La présente question a été ajoutée par la ville de Rock Forest. Elle avait pour objectif de connaître les raisons qui motivent les gens à prendre le transport en commun comme moyen de locomotion. Une question à choix multiples présentant six modalités non lues a donc été incluse dans le questionnaire.

Q6. Généralement, pour votre famille, pour quelle(s) raison(s) prenez-vous ou allez-vous prendre (prendriez-vous) l'autobus?

(Pour l'enquêteur, ne pas lire les raisons)

- le travail
- les études
- le magasinage
- les loisirs
- par affaire
- autres :

La formulation de la question respecte les principes de rédaction de base. Soulignons que cette question à choix multiples présente la particularité que les modalités ne sont pas lues à l'interrogé. Ce type de question remplace en fait les questions subjectives à développement court. Elle possède l'avantage d'éliminer le traitement des phrases construites par le répondant. L'enquêteur écoute et note dans les modalités celles qui rejoignent le plus les dires du répondant. Bien que cette question élimine le temps de traitement associé aux questions subjectives, elle est tout de même soumise à l'aléa de la compréhension de l'enquêteur.

L'interaction entre cette question et la population fut assez bonne, mais l'ensemble des modalités était trop large. En effet, pour plusieurs familles, il est arrivé que cinq des six modalités furent cochées en même temps. A posteriori, cette question ne soulève pas grand intérêt et rallonge inutilement le questionnaire. Cette question serait donc éliminée dans une autre étude similaire.

5.5.1.7 La question #7

Le prochain objectif consiste à faire resssortir l'opinion des gens en ce qui concerne les trajets du réseau de transport en commun sillonnant la ville de Rock Forest. Pour être capable d'obtenir l'information, une question à choix multiples offrant six modalités se révèle être un outil classique et efficace.

Q7. Présentement, pour votre famille, diriez-vous que dans leur ensemble les trajets du réseau de transport en commun pour la ville de Rock Forest sont

- excellents
- très bons
- bons
- mauvais
- très mauvais

(non lue) - pas d'opinion

La formulation de cette question est classique et respecte toutes les règles de rédaction usuelles. Elle possède la particularité que seule les cinq premières modalités sont lues par le

sondeur. Ceci a pour conséquence d'éviter que le répondant se laisse emporter par un élan de paresse et réponde trop facilement qu'il n'a pas d'opinion.

L'interaction entre cette question et la population fut relativement bonne. En fait, l'enquêteur se devait de souligner chacun des mots (notamment le mot trajet) de la question. En effet, dans la strate rurale, le niveau de frustration des gens face au transport collectif était assez grand et il est possible que cette frustration se soit transposée sur la réponse à cette question.

5.5.1.8 La question #8

L'objectif de la présente question est de connaître le degré de satisfaction des usagers et d'éventuels usagers face au transport en commun. Tout comme la précédente question, nous avons opté pour une question à choix multiples présentant six modalités dont la dernière est non lue.

Q8. Présentement, pour votre famille, diriez-vous que votre niveau de satisfaction, en ce qui concerne le transport en commun de la ville de Rock Forest, est

- très élevé
- élevé
- bon
- peu élevé
- très peu élevé

(non lue) - pas d'opinion

Tout comme la question précédente, la formulation classique de cette question respecte tous les standards de rédaction relatifs à ce type de question. L'interaction entre cette question et la population fut parmi les meilleures du questionnaire. En effet, contrairement à la question précédente, aucune fois l'enquêteur n'a eu à insister sur un mot de la phrase plus qu'un autre.

A posteriori, les questions à choix multiples s'avèrent être très efficaces. Une seule ombre au tableau se dessine; la relation entre les modalités est une relation d'ordre. Les outils servant au traitement de l'information s'avèrent moins performants. En effet, à quel point la modalité "bon" est loin de la modalité adjacente "peu élevé"? Ainsi, dans une prochaine étude une échelle d'appréciation de type graphique serait possiblement un meilleur item.

5.5.1.9 La question #9

L'objectif de cette dernière question consiste à recueillir toutes les suggestions susceptibles d'améliorer le transport en commun. Comme il était impossible d'élaborer l'ensemble des réponses possibles, une question à interprétation subjective à réponse limitée a été conçue.

Q9. À propos du transport en commun pour la ville de Rock Forest, avez-vous des suggestions concernant de nouvelles destinations, de nouveaux trajets ou de nouveaux horaires?

_____

_____

_____

_____

La formulation de la question a été sujette à plusieurs modifications. En effet, lors de l'étude préliminaire le sondeur s'est rendu compte que les gens avaient peu de suggestions à donner. Ainsi, nous avons reformulé la question afin d'amener les répondants sur des pistes de réflexion. Par exemple pour guider les gens, nous avons ajouté <<concernant de nouvelles destinations, de nouveaux trajets ou de nouveaux horaires?>>. Est-ce que cette tentative n'a pas guidé l'opinion des gens? Peut être, mais vaut-il mieux avoir de l'information sur des domaines précis que pas du tout? Poser la question, c'est y répondre.

Finalement, l'interaction entre la question et la population fut très bonne. Son emplacement était optimal et la durée de remplissage de cette question a été relativement courte. A posteriori, la suggestion <<nouveau point de correspondance>> serait ajoutée à l'ensemble des suggestions que propose la question.

5.5.2 Le questionnaire concernant la bibliothèque

Conjointement à l'étude sur le transport en commun se retrouve en second plan l'étude sur la bibliothèque municipale. Elle avait pour but principal de répondre aux trois questions suivantes : <<Pourquoi les gens ne vont pas à la bibliothèque?>>, <<Est-ce parce que les

services ne sont pas bons?>> et <<Quel serait l'opinion des gens dans l'éventualité d'un déménagement de la bibliothèque?>>.

Pour atteindre ces objectifs, sept questions ont été posées. Nous allons jeter un regard attentif sur chacune d'entre elles. Comme à la section précédente, nous allons nous efforcer de répondre aux questions suivantes : <<Quelle est l'objectif que doit poursuivre la question courante?>>, <<Quel est le type de question qui permettrait d'obtenir l'information voulue?>>, <<La formulation de la question satisfait-elle les principes de rédaction?>>, <<Quelle a été l'interaction entre les gens et la question lors de la pré-enquête ou lors de l'enquête?>> et finalement, <<A posteriori, quels seraient les correctifs à apporter?>>.

5.5.2.1 La question #1

L'objectif premier de cette première question a pour but de savoir si les gens connaissent simplement où se trouve la bibliothèque de Rock Forest. En effet, il est possible que des gens ne se présentent pas à la bibliothèque simplement en raison du fait qu'il ne savent pas où elle se trouve. L'univers des réponses est simple, on le sait ou on ne le sait pas. Une question à alternative a donc été élaborée.

Q1. Est-ce que vous ou quelqu'un de votre famille savez où se trouve la bibliothèque municipale de Rock Forest?

- oui Passer à la question 2
- non Passer à la question 6

Comme à l'habitude, la formulation de cette question respecte les règles de rédaction. Cependant mentionnons que la modalité, non, de cette question n'a pas vraiment été mise à l'épreuve lors de l'étude préliminaire. En effet, celle-ci a été menée, volontairement, près de la bibliothèque.

L'interaction entre la question et la population fut très bonne. Cependant après qu'un répondant ait répondu non à la question, le sondeur a noté qu'après quelques minutes de discussion, plusieurs d'entre eux disaient se souvenir soudainement de l'emplacement de la bibliothèque. Ces personnes ont été tout de même inscrites sous la modalité "non". En effet, s'ils ne se souviennent pas de l'emplacement de la bibliothèque, cela veut dire que la

bibliothèque ainsi que les services offerts ne sont pas suffisamment bien publicisés. Mentionnons que si l'interrogé répondait non à cette question, l'enquêteur passait immédiatement à la question 6.

5.5.2.2 La question #2

L'objectif de cette question est de savoir si la famille à laquelle l'on s'adresse utilise ou non la bibliothèque. Comme pour la question précédente, l'univers des réponses étant de nature dichotomique, une question à alternatives s'avère le meilleur des choix.

Q2. Présentement, est-ce que vous ou quelqu'un de votre famille êtes des usagers de la bibliothèque?

- oui Passer à la question 3
- non Passer à la question 6

La formulation de la question ne vise qu'un seul objectif et de plus elle est concise, claire, et rapide. Elle respecte bien les principes de rédaction de base relatifs à ce type de question. Aucune fois le sondeur n'a eu à préciser quel que détail que ce soit. L'interaction entre cette question et la population a été très bonne.

Cependant, notez que cette question s'est possiblement avérée biaisée par la saison des vacances d'été. Selon les résultats obtenus de la question suivante, il est clair que l'été représente une saison morte pour la bibliothèque de Rock Forest. A posteriori, cette étude aurait dû être menée lors d'une autre saison que celle de l'été. De plus, l'étude préliminaire aurait dû être menée ailleurs que près de l'emplacement de la bibliothèque. Ce qui nous aurait peut-être permis de voir venir le phénomène au lieu de le subir.

5.5.2.3 La question #3

L'objectif relatif à cette troisième question est de dégager une estimation de l'affluence au comptoir de la bibliothèque. Nous avons défini un univers de réponses qui nous semblait a priori très satisfaisant. Une question à choix de réponses nous a semblé être l'outil le plus adéquat pour faire ressortir l'information espérée.

Q3. Si oui, combien de fois y êtes-vous allés le mois dernier?

- 0
- 1
- 2
- 3 et plus

La formulation de la question ne renfermait aucune ambiguïté et respectait selon nous les normes de rédaction. Seulement, comme nous venons de le mentionner, l'affluence au comptoir est saisonnier, ainsi pratiquement toutes les réponses ont été 0. Comme cette question n'a pas retourné l'information attendue, aucune analyse, sinon le tableau des fréquences, n'a pu être exécutée.

L'interaction fut bonne mais les réponses furent peu significatives face à l'information que nous voulions retirer. Il faut avouer que nous ignorions jusqu'à quel point la saison estivale était tranquille pour une bibliothèque. Ainsi, a posteriori, nous allons, avant de rédiger le questionnaire, nous informer auprès de l'organisme impliqué du caractère saisonnier de ses affluences.

5.5.2.4 La question #4

L'objectif poursuivi par la présente question consiste à mesurer la qualité des services offerts par la bibliothèque. Pour être capable d'obtenir l'information, une question à choix multiples offrant six modalités a été choisie.

Q4. Diriez-vous que la qualité des services offerts par la bibliothèque est

- excellente
- très bonne
- bonne
- mauvaise
- très mauvaise

(non lue) - pas d'opinion

Comme les autres questions à choix multiples, la formulation de celle-ci est assez standard. Les modalités s'approchant le plus possible des caractéristiques de l'opinion cherchée

simplifient et améliorent grandement la compréhension de l'enquêté. L'interaction fut très bonne et par chance les réponses étaient indépendantes de la saison estivale. A posteriori, nous dirions que ce type de question est stable au niveau de l'interaction entre l'enquêté et l'enquêteur, qualité qui s'avère être très importante dans un questionnaire.

5.5.2.5 La question #5

L'objectif poursuivi par cette cinquième question consiste à savoir si les gens ne fréquentent pas à la bibliothèque en raison de la mauvaise sélection de livres. Tout comme la précédente question nous avons opté pour une question à choix multiples. Cependant, afin de se rapprocher le plus possible des caractéristiques de l'opinion recherchée, elle ne contient que quatre modalités.

Q5. Diriez-vous que la variété ou la sélection des livres est

- très satisfaisante
- satisfaisante
- peu satisfaisante

(non lue) - pas d'opinion

Trois modalités furent suffisantes pour exprimer l'ensemble des commentaires face à la sélection des livres. Cette question satisfait tout de même les principes de rédaction. Les modalités sont par leur forme très suggestives. L'interaction entre les répondants et le sondeur fut très bonne. A posteriori, aucun changement ne serait apporté à cette question.

5.5.2.6 La question #6

L'objectif que poursuit cette question formalise un des principaux buts de la seconde partie de ce questionnaire. L'étude des réponses à cette question va nous permettre de connaître l'opinion des gens en ce qui concerne l'éventualité d'un déménagement. Une question à choix multiples fut donc proposée.

Q6 Que pensez-vous de l'éventualité d'un déménagement de la bibliothèque aux Terrasses Rock Forest? Êtes-vous

- tout à fait d'accord
- d'accord
- pas du tout d'accord

(non lue) - pas d'opinion

La formulation de la question était claire et précise. Peu de gens sans opinion se sont manifestés. Cependant, le sondeur a dû expliquer préalablement où se trouvait l'emplacement présent de la bibliothèque et où allait possiblement se trouver la bibliothèque si elle déménageait. Mais cela n'a pas posé de problème majeur. A posteriori, une carte des lieux serait indispensable au sondeur. En effet, certains citoyens sont vraisemblablement plus visuels qu'auditifs.

5.5.2.7 La question #7

Cette dernière question a comme objectif de recueillir les suggestions des citoyens de la ville de Rock Forest. Ainsi, une question à interprétation subjective à réponses limitées a été développée.

Q7. Avez-vous des suggestions?

_____

Vous noterez qu'une seule ligne délimitant les réponses a été inscrite. Ceci ne respecte pas un des principes de rédaction de ce type de question. Cependant, il ne restait plus de place dans le bas du questionnaire et imprimer une feuille pour n'inscrire que cinq lignes aurait entraîné des frais supplémentaires inutiles.

Dans l'ensemble les gens se sont sentis concernés par la démarche et le projet de déménagement. La population a été très volubile face à cette question que ce soit pour des suggestions d'achat de livres ou pour des idées de rénovation de la bibliothèque. Les résultats de cette question ont permis de recueillir une liste de suggestions particulièrement intéressantes et inédites. A posteriori, ce type de question termine très bien un questionnaire et les suggestions recueillies sont d'une grande richesse.

Chapitre 6

Les données

Ce chapitre discute des différentes étapes qui suivent la conception du questionnaire et le tirage de l'échantillon. En effet, nous y discutons des différentes méthodes existantes permettant la cueillette des données. Nous y abordons ensuite les étapes de la codification, de la saisie des données, du contrôle de la qualité de ces dernières. Finalement, nous traitons de l'épineux problème du traitement de la non-réponse.

6.1 La collecte des données

Dans cette section nous discutons des principales méthodes de collecte de données. Plusieurs modes de collecte existent et peuvent être utilisés. On peut faire des interviews directs, des enquêtes par téléphone, des enquêtes postales, ou une combinaison de ces dernières. L'important est de choisir celle qui s'adapte le mieux à notre étude. Il faut bien comprendre qu'un mode qui fonctionne bien pour une enquête peut être catastrophique pour une autre.

Chaque fois que le budget le permet, on préfère le contact enquêteur-enquêté. Les erreurs d'observations sont moindres et les taux de réponses sont plus élevés. Cependant, parmi les dominantes des erreurs de mesure, on peut toujours mettre en évidence un effet enquêteur assez fort : l'âge de l'enquêteur, son charisme, son approche ne sont pas sans conséquence sur la qualité des réponses obtenues. On peut amoindrir cette source d'erreur d'observation en offrant aux enquêteurs un bon encadrement et des périodes de formation avant chacune des collectes. Dans ARDILLY [1], l'auteur mentionne qu'il est important d'effectuer des enquêtes de contrôle des enquêteurs, de relire les fichiers de réponses, de recontacter certains enquêtés dans le cas où les réponses sont aberrantes, etc.

Au Canada, les enquêtes par téléphone ne sont pas rares. Ce type de collecte de données possède l'avantage de minimiser les coûts de déplacements et le temps global de l'enquête. Cependant Ningay et Greenwell ont montré dans leur étude (Journal of Official Statistic 1989) que l'ordre de lecture des modalités d'une question par téléphone a une forte influence

auprès de la personne contactée. Ils soulignent que celle-ci retient, lorsqu'on énumère plusieurs possibilités, en priorité la première modalité qui lui semble compatible avec ce qu'elle ressent sans demander à l'enquêteur de répéter les divers choix possibles. De plus, il est difficile ou impossible d'expliquer le sens de la question aux répondants. Mentionnons finalement que cette méthode peut s'avérer très intéressante dans les études compromettantes pour l'enquêté, car elle présente l'avantage de sauvegarder partiellement l'anonymat.

Lorsque le budget est plus restreint, certaines enquêtes, simples et rapides, comme les enquêtes accessoires, se font par voie postale. Les taux de réponses sont en général peu élevés. Les taux varient dans les bonnes études entre soixante et soixante-quinze pour cent. En bref, cette méthode de collecte souffre du syndrôme de l'oubli chez les enquêtés; la lettre de l'enquête se retrouve souvent sous la pile de courrier pour finir ses jours dans la poubelle. Certains artifices peuvent cependant être utilisés pour améliorer substantiellement le taux de réponse. Par exemple le responsable de l'enquête peut glisser dans l'enveloppe d'envoi une enveloppe de retour préadressée et affranchie, glisser une pièce de monnaie dans l'enveloppe, joindre des articles de promotion, promettre des cadeaux seulement si le retour est effectué, etc. Les résultats issus de cette méthode de collecte sont aussi très sensibles aux questions ambigües, mal définies ou mal formulées. Il est en effet impossible de savoir si les questions ont été bien comprises ou bien interprétées. En général, cette méthode est utilisée pour obtenir une certaine quantité de réponses à moindre coût pour ensuite être jumelée à une relance soit par l'entremise d'entrevue téléphonique ou par l'entremise d'entrevue personnelle.

En ce qui concerne notre étude, le budget disponible a permis au responsable de l'enquête d'effectuer une collecte des données par l'entremise d'interviews à domicile. Le taux de réponse a été de 92%. Mentionnons que les vacances estivales ont été la cause principale de la non-réponse. Dans la section 6.3 nous allons discuter de la méthode d'imputation des données qui a été utilisée pour contrebalancer la non-réponse.

6.2 La codification, la saisie et la vérification des données

La collecte des données doit être suivie d'une phase de contrôle permettant de tester la conformité du comportement des enquêteurs aux instructions qu'ils ont reçues. Un fois cette opération effectuée, il reste à les transformer pour permettre leur passage sur une base de données. Ceci se fait généralement en deux étapes: premièrement le codage, où les réponses

numériques ou littérales sont traduites selon un code, et deuxièmement la saisie des données, où l'information codée est placée sur un support informatique. Simultanément sont effectués divers contrôles de qualité afin de repérer les erreurs de mesure.

L'opération de chiffrement s'appuie en général sur des nomenclatures donnant pour un caractère qualitatif donné un code correspondant à chacune de ses modalités. Par exemple pour chiffrer une question dont les modalités sont des langues étrangères, on peut procéder ainsi: Anglais=01, Espagnol=02, Italien=03, etc. La représentation disjonctive complète des données est une autre technique de chiffrement fort utilisée lorsque les variables qualitatives en présence offrent plusieurs modalités. En bref, la forme disjonctive d'une variable associe à chacune des modalités une fonction indicatrice faisant foi de la présence ou de l'absence de chacune des modalités chez le répondant. La difficulté liée à cette opération est intimement liée aux réponses inattendues qui sont difficiles à inclure dans une classe. En principe les données numériques n'ont pas besoin d'être codées. Signalons que, malgré tout, certaines simplifications peuvent être faites au moment du chiffrement et que les réponses numériques peuvent être remplacées par des données qualitatives. Par exemple, si une question demande l'âge des individus, on peut être tenté de regrouper les individus en classe d'âge, ce qui nécessite un code. La saisie des données est l'étape suivante qui se doit d'être effectuée avec la plus grande attention.

Après avoir réalisé la saisie, il est préférable d'effectuer quelques vérifications dans le but d'éliminer un maximum d'erreurs d'observation et de saisie. Les pratiques employées sont assez diverses. Une première possibilité consiste à effectuer un contrôle comptable en effectuant une vérification des additions, des pourcentages, des distributions marginales et conditionnelles dans le cas de tableaux à plusieurs entrées. Une seconde possibilité consiste à repérer les réponses ou résultats qui paraissent incompatibles. Ce type de redressement repose cependant sur des bases subjectives et ne doit être fait que par un personnel qualifié. La venue de l'informatique nous amène un troisième type de pratique qui, dès la saisie, permet de contrôler la qualité intrinsèque des données. En effet, il est maintenant possible de programmer des messages d'avertissement soulignant les principales contradictions entre les diverses variables. Par exemple, si l'âge de l'individu est de 15 ans, il est impossible qu'il puisse se classer parmi les retraités. Enfin, la solution optimale consiste à contrôler l'ensemble des variables en tirant un échantillon de questionnaires; soit que l'on recodifie et

ressaisit les données et que l'on confronte informatiquement à la saisie de base, soit qu'on contrôle visuellement la validité de la codification et de la saisie à l'écran.

6.3 Le traitement de la non-réponse

Malgré une collecte des plus efficaces, il est rare que l'échantillon, initialement prévu de taille n, puisse produire exactement n résultats exploitables. On dit qu'il y a non-réponse vis-à-vis la variable Y pour l'individu échantillonné i dès lors qu'on ne dispose pas de la valeur Y_i relative à cet individu. On distingue deux types de non-réponse. Il y a la non-réponse totale où l'unité échantillonnée ne répond à aucune question; elle est généralement causée par un refus, par une absence ou parce que l'enquêté a perdu le questionnaire. Et il y a la non-réponse partielle où l'unité échantillonnée refuse de répondre à certaines questions qu'elle trouve trop indiscrètes ou tout simplement dont elle ne connaît pas la réponse. Notez que le non-réponse partielle peut être introduite lors de la phase de saisie et de codification (valeur aberrante, problème informatique, etc.). Quel que soit le type de non-réponse, il en résulte pour les estimateurs, l'introduction d'un biais et une diminution de la précision. Pour traiter la non-réponse, on adopte des méthodes qui font appel soit à la repondération des individus répondants, soit à l'imputation de valeurs aux individus non-répondants.

Le principe de la méthode de repondération est simple. On cherche à se rapprocher au maximum de l'estimateur "utopique" et idéal en ne considérant que les individus répondants. Pour y arriver, on modifie leurs poids de sondage initiaux pour tenir compte de la non-réponse. Cette technique repose hélas sur une hypothèse discutable qui stipule que chacun des non-répondants "aurait pu" répondre de la même façon qu'un répondant. Ce qui nous amène à accepter qu'une composante du biais risque d'être importante. En effet, il n'est pas évident d'affirmer que le comportement des non-répondants n'est en fait qu'une extension du comportement des répondants.

Le principe des méthodes d'imputation est une idée plus naturelle. Elle consiste à estimer la valeur inconnue de la variable d'intérêt Y_i pour chaque individu i non-répondant par une valeur y_i^* pertinente. Différentes méthodes peuvent être utilisées. En voici quelques-unes : on peut effectuer le retrait de tous les relevés ayant des données manquantes ou incomplètes. Cette méthode très simple possède les inconvénients que la taille de l'échantillon peut être très réduite et que si les individus enlevés sont très différents, les estimés obtenus peuvent

être fortement biaisés. On peut imputer les données manquantes à l'aide de la moyenne. Cette technique, possiblement la plus utilisée possède l'inconvénient de réduire trompeusement les variances calculées. On peut utiliser une des deux méthodes du donneur (hot deck et cold deck). La méthode hot deck consiste à trouver des donneurs à même le fichier initial, qui sont en fait des individus ayant des registres complets, et à imputer les données manquantes à l'aide des données semblables chez les donneurs. La méthode cold deck véhicule la même idée à la différence que les donneurs sont pris à l'extérieur du fichier initial, soit à l'aide des études préliminaires ou soit par l'entremise d'une étude comparable (on peut utiliser ici les techniques de régression). On peut aussi imputer les données manquantes à partir de donneurs sélectionnés au hasard. Finalement, on peut imputer les données manquantes à l'aide de déductions subjectives. Le lecteur intéressé trouvera dans RUBIN [6] un traitement très complet de la non-réponse.

Dans notre étude, nous avons utilisé deux méthodes différentes pour le traitement des données manquantes. Premièrement nous avons enlevé les individus dont les adresses n'existaient plus (maisons détruite ou abandonnée). Deuxièmement, nous avons utilisé un cold deck à l'aide du fichier des données préliminaires.

Chapitre 7

L'analyse des données

Après le recueil des données, la démarche statistique consiste à traiter et interpréter les informations recueillies. Elle comporte deux grands aspects : l'aspect descriptif et l'aspect déductif. Dans ce chapitre, nous allons discuter des méthodes couvrant ces deux aspects qui ont été utilisées dans cette étude.

7.1 La statistique descriptive

Le but de la statistique descriptive est de synthétiser, résumer, structurer l'information contenue dans les données. Elle utilise pour cela des représentations des données sous forme de tableaux, de graphiques et d'indicateurs numériques. Le rôle de la statistique descriptive est de mettre en évidence les propriétés de l'échantillon et de suggérer des hypothèses.

Les minimums, les maximums, les tableaux des fréquences, les diagrammes en bâtons et les tableaux à doubles entrées sont les principaux outils utilisés lorsque les données sont qualitatives ou quantitatives discrètes. Lorsqu'on a des variables multidimensionnelles, on peut aussi utiliser l'analyse des correspondances. Les points d'inflexions, les effectifs de classes, les histogrammes et les courbes de concentration ne sont que quelques exemples d'outils utilisés pour visualiser les données lorsque celles-ci sont quantitatives et continues. Dans le cas multidimensionnel, on peut recourir à l'analyse en composantes principales. Lorsque les données sont numériques, on peut extraire des indicateurs numériques. Ils tentent en général de résumer une série d'observations par l'entremise d'une seule valeur numérique. Il est cependant insuffisant de résumer une série d'observations par un seul indicateur.

7.2 L'inférence statistique

Un premier but de l'inférence statistique est d'étendre les propriétés constatées sur l'échantillon à la population toute entière. Par exemple, nous pouvons estimer un total, une

moyenne ou une proportion. Nous pouvons aussi calculer les intervalles de confiance associés. Elle consiste aussi à valider ou infirmer des hypothèses a priori formulées après la phase descriptive.

Au chapitre 3, lorsque nous avons fait le choix du plan d'échantillonnage de type stratifié, nous avons par le même coup déterminé les expressions de l'estimateur pour la proportion et de l'intervalle de confiance associé. Ce dernier a été utilisé sous la forme présentée au chapitre 3. Cependant dans l'élaboration de notre étude pratique, nous avons utilisé le même estimateur pour la proportion que nous avons dû associer à un autre intervalle de confiance.

Dans plusieurs études, la strate à laquelle appartient un individu ne sera connue qu'après que les données aient été collectées. Par exemple le sexe, l'âge ou le niveau d'éducation peuvent être connus par les données officielles, mais selon lesquels la stratification peut être difficile à réaliser a priori. Un plan d'échantillonnage qui consiste à sélectionner des individus selon un plan aléatoire simple mais à appliquer les procédures d'estimation du plan aléatoire stratifié est appelé un plan de post-stratification. Les individus ne sont classés dans les strates qu'a posteriori.

Comme l'expression de l'estimateur de la proportion pour les plans stratifié et post-stratifié est le même, nous allons seulement présenter les fondements théoriques permettant de dégager l'intervalle de confiance pour le paramètre de la proportion appartenant au plan de post-stratification. Pour mieux illustrer notre propos, nous présentons d'abord le plan aléatoire de Bernoulli. Ce plan, bien que similaire au plan aléatoire simple, se distingue par le fait que la taille de l'échantillon est tirée de façon aléatoire. C'est par l'entremise de ce plan que nous allons mettre en évidence la théorie nécessaire pour dégager l'intervalle de confiance cherché.

7.2.1 Le plan aléatoire de Bernoulli

Nous allons ici aborder la méthode de tirage dite de Bernoulli. Cette méthode d'échantillonnage est similaire à celle du plan aléatoire simple à la différence près que la taille de l'échantillon à tirer est aléatoire. Il s'agit d'un plan d'échantillonnage de type conditionnel à la taille de l'échantillon.

En théorie, le plan de Bernoulli repose sur l'hypothèse qu'on peut admettre n'importe quelle taille d'échantillon comprise entre 0 et N. La taille de l'échantillon est donc une variable aléatoire qui admet une loi de probabilité. Lorsque cette loi de probabilité est une binomiale , on dit qu'on a un plan de Bernoulli. De cette constatation, il vient naturellement que la variable aléatoire, n_S suit une loi binomiale de paramètres N et $p = {E(n_s)}/{N}$.

L'estimateur de la proportion que propose cette méthode est le suivant :

$$p_{ys} = \begin{cases} \frac{1}{n_S}\sum_{i=1}^{n_S} y_i & \text{si } n_S \geq 1 \\ 0 & \text{sinon} \end{cases} \quad \text{où } y_i = \begin{cases} 1 \text{ si l' individu i possède le caractère} \\ 0 \text{ sinon} \end{cases}$$

Lorsque $n_S \geq 1$, cet estimateur a la même forme que celui du plan aléatoire simple. Définissons Ai l'événement $n_S \geq i$.

$$\mathrm{P}(A1) = 1 - \mathrm{P}(A1^c) = 1 - \mathrm{P}(n_S = 0) = 1 - (1-p)^N \cong 1 - e^{-E(n_S)/p},$$

où $n = \mathrm{E}(n_S) = Np$ est la taille d'échantillon espérée. On remarque que même pour une petite taille d'échantillon espérée, l'événement A1 est presque sûr de survenir.

THÉORÈME 7.2.1

Conditionnellement à A1 et à n_S , l'estimateur sans biais de la proportion pour le plan de Bernoulli est

$$p_{ys} = \begin{cases} \frac{1}{n_S}\sum_{i=1}^{n_S} y_i & \text{si } n_S \geq 1 \\ 0 & \text{sinon} \end{cases} \quad \text{où } y_i = \begin{cases} 1 \text{ si l' individu i possède le caractère} \\ 0 \text{ sinon} \end{cases}$$

De plus, sa variance théorique conditionnelle à A1 et à n_S s'écrit de la façon suivante :

$$Var\left(p_{\mathrm{ys}} \middle| n_S, \mathrm{A1}\right) = \frac{P_y\left(1-P_y\right)}{n_s}\left(\frac{N-n_s}{N-1}\right) = S_y^2\left(\frac{1}{n_S} - \frac{1}{N}\right).$$

Où $S_y^2 = \frac{N}{N-1}\left[P_y\left(1-P_y\right)\right]$ et où P_y est la proportion de la population à estimer.

Démonstration :

Montrons premièrement que cet estimateur est sans biais. Il suffit de montrer que

$$\mathrm{E}\left(p_{ys} \mid n_S \text{ et } A1\right) = P_y.$$

Comme nous l'avons déjà fait à plusieurs reprises, définissons l'indicateur suivant :

$$I_i = \begin{cases} 1 \text{ si l' individu i est dans l' échantillon} \\ 0 \text{ sinon} \end{cases}.$$

Comme $n_S \geq 1$, nous sommes dans la même situation où nous avons un P.A.S. avec $n = n_s$. Et nous avons vu que dans ces conditions, $\mathrm{P}(I_i = 1) = \frac{n_s}{N} \quad \forall i = 1,2,\ldots,N$.

On peut alors écrire :

$$\begin{aligned} \mathrm{E}\left(p_{ys} \mid n_S \text{ et } A1\right) &= \mathrm{E}\left(\frac{1}{n_s}\sum_{i=1}^{n_s} y_i \mid n_S \text{ et } n_S \geq 1\right) \\ &= \frac{1}{n_s}\sum_{i=1}^{N} Y_i \mathrm{E}(I_i) \\ &= \frac{1}{n_s}\sum_{i=1}^{N} Y_i\left(\frac{n_s}{N}\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} Y_i \\ &= P_y \end{aligned}$$

La variance de l'estimateur se calcule de façon similaire à celle du plan aléatoire simple:

$$Var\left(p_{ys} \mid n_s \text{ et } A1\right) = \frac{1}{n_s^{\,2}}\left[\sum_{i=1}^{N} Y_i^2 Var(I_i) + \Sigma_{i \neq j} \Sigma Y_i Y_j \operatorname{cov}(I_i, I_j)\right].$$

D'autre part, conditionnellement à A1 et n_s, on trouve que

$$\operatorname{cov}(I_i, I_j) = \mathrm{E}(I_i I_j) - \mathrm{E}(I_i)\mathrm{E}(I_j) = \frac{n_s(n_s-1)}{N(N-1)} - \left(\frac{n_s}{N}\right)^2.$$

Ce qui nous permet d'écrire que:

$$\begin{aligned}
Var\left(p_{ys} \mid n_s \text{ et } A1\right) &= \frac{1}{n_s^2}\left[\frac{n_s}{N}\left(1-\frac{n_s}{N}\right)\sum_{i=1}^{N} Y_i^2 + \left[\frac{n_s}{N}\left(\frac{n_s-1}{N-1}\right) - \left(\frac{n_s}{N}\right)^2\right]\Sigma_{i\neq j}\Sigma Y_i Y_j\right] \\
&= \frac{1}{n_s}\left(\frac{N-n_s}{N-1}\right)\left[\frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \frac{1}{N^2}\left[\sum_{i=1}^{N} Y_i^2 + \Sigma_{i\neq j}\Sigma Y_i Y_j\right]\right] \\
&= \frac{1}{n_s}\left(\frac{N-n_s}{N-1}\right)\left[\frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \frac{1}{N^2}\left[\sum_{i=1}^{N} Y_i\right]^2\right] \\
&= \frac{1}{n_s}\left(\frac{N-n_s}{N-1}\right)\left[P_y - P_y^2\right] \\
&= \frac{1}{n_s}\left(\frac{N-n_s}{N}\right)S_y^2 \quad \text{où } S_y^2 = \frac{N}{N-1}P_y\left(1-P_y\right) \\
&= S_y^2\left(\frac{1}{n_s} - \frac{1}{N}\right)
\end{aligned}$$

■

Il est à noter que cette démonstration est valide quelle que soit la distribution de probabilité de n_s.

THÉORÈME 7.2.2

L'estimateur sans biais de la variance théorique n'est défini que pour $n_s \geq 2$ et il s'écrit:

$$\hat{V}ar\left(p_{ys} \mid n_s \text{ et } A2\right) = s_y^2\left(\frac{1}{n_s} - \frac{1}{N}\right).$$

Où $s_y^2 = \frac{n_s}{n_s-1}p_y\left(1-p_y\right)$.

Démonstration :

On sait que A2 représente l'événement $n_s \geq 2$. Montrons que cet estimateur est un estimateur sans biais de la variance théorique conditionnelle de p_{ys}. Calculons donc son espérance :

$$\mathrm{E}\left(\hat{V}ar\left(p_{ys}\,|\mathrm{n_s}\text{ et }A2\right)\right) = \mathrm{E}\left(s_y^2\left(\frac{1}{n_s}-\frac{1}{N}\right)\Big|\mathrm{n_s}\text{ et }A2\right)$$

$$= \left(\frac{1}{n_s}-\frac{1}{N}\right)\mathrm{E}\left(s_y^2|\mathrm{n_s}\text{ et }A2\right)$$

Les conditions $\mathrm{n_s}$ et $A2$ nous ramènent au cas d' un P.A.S de taille $\mathrm{n_s}$ fixée. On peut donc écrire $\mathrm{E}\left(s_y^2|\mathrm{n_s}\text{ et }A2\right) = \mathrm{E}\left(s_y^2\right) = S_y^2$

$$= \left(\frac{1}{n_s}-\frac{1}{N}\right)S_y^2$$

$$= Var\left(p_{ys}\,|\mathrm{n_s}\text{ et }A2\right)$$

■

Il est à noter que cette démonstration est valide quelle que soit la distribution de probabilité de n_s. Ces deux dernières démonstrations sont en fait des variantes des théorèmes 3.1.2 et 3.1.3 si l'on prend le soin de noter que les conditions sur la variable aléatoire n_s nous ramènent à un plan aléatoire simple.

Les moments conditionnels théoriques que nous venons de développer pour le plan de Bernoulli peuvent être utilisés afin d'en déduire les moments non conditionnels correspondants. Posons $E_1(\,)$ et $V_1(\,)$ l'espérance et la variance non conditionnelle en fonction de la loi de n_s que voici :

$$\mathrm{P}(n_s = j) = \binom{N}{j}\frac{p^j(1-p)^{N-j}}{\mathrm{P}(A1)};\quad j = 1,2,\ldots,N$$

Il s'agit d'une loi binomiale de paramètres N et p tronquée à zéro.

THÉORÈME 7.2.3

Soit l'estimateur de la proportion pour le plan de Bernoulli,

$$p_{ys} = \begin{cases} \dfrac{1}{n_s}\sum_{i=1}^{n_s} y_i & \text{si } n_s \geq 1 \\ 0 & \text{sinon} \end{cases} \qquad \text{où } y_i = \begin{cases} 1 \text{ si l' individu i possède le caractère} \\ 0 \text{ sinon} \end{cases}$$

L'espérance non conditionnelle de celui-ci est égale à $P(A1)P_y$. De plus, sa variance théorique non conditionnelle s'écrit de la façon suivante :

$$Var(p_{ys}) = \mathrm{P}(A1)(1-\mathrm{P}(A1))P_y^2 + \mathrm{P}(A1)\left[\mathrm{E}_1\left(\frac{1}{n_s}\right) - \frac{1}{N}\right]S_y^2 \ .$$

Où $S_y^2 = \dfrac{N}{N-1}\left[P_y(1-P_y)\right]$.

Démonstration :

Développons d'abord l'espérance non conditionnelle de cet estimateur :

$$\begin{aligned} \mathrm{E}(p_{ys}) &= (1-\mathrm{P}(A1))\mathrm{E}(p_{ys} \mid A1^c) + \mathrm{P}(A1)\mathrm{E}(p_{ys} \mid A1) \\ &= \mathrm{P}(A1^c)(0) + \mathrm{P}(A1)\mathrm{E}_1\left[\mathrm{E}(p_{ys} \mid \mathrm{n_s \text{ et } A1})\right] \\ &= \mathrm{P}(A1)\mathrm{E}_1\left[P_y\right] \quad \text{car } \mathrm{E}(p_{ys} \mid \mathrm{n_s \text{ et } A1}) = P_y \\ &= \mathrm{P}(A1)P_y \end{aligned}$$

Afin de calculer la variance non conditionnelle cherchée, développons dans un premier temps l'erreur quadratique moyenne de notre estimateur :

$$\begin{aligned} EQM(p_{ys}) &= \mathrm{E}\left[\left(p_{ys} - P_y\right)^2\right] \\ &= (1-\mathrm{P}(A1))\mathrm{E}\left[\left(p_{ys} - P_y\right)^2 \mid \mathrm{A1}^c\right] + \mathrm{P}(A1)\mathrm{E}\left[\left(p_{ys} - P_y\right)^2 \mid \mathrm{A1}\right] \end{aligned}$$

$$EQM(p_{ys}) = \mathrm{P}(A1^c)\mathrm{E}\left[\left(P_y\right)^2\right] + \mathrm{P}(A1)\mathrm{E}_1\left[\mathrm{E}\left[\left(p_{ys} - P_y\right)^2 \,\middle|\, \text{A1 et n}_s\right]\right]$$

$$= \mathrm{P}(A1^c)P_y^{\,2} + \mathrm{P}(A1)\mathrm{E}_1\left[\mathrm{E}\left[\left(p_{ys} - P_y\right)^2 \,\middle|\, \text{A1 et n}_s\right]\right]$$

$$= \mathrm{P}(A1^c)P_y^2 + \mathrm{P}(A1)\mathrm{E}_1\left[Var\left[p_{ys} \,\middle|\, \text{A1 et n}_s\right]\right]$$

$$= \mathrm{P}(A1^c)P_y^2 + \mathrm{P}(A1)\mathrm{E}_1\left[\left[\left(\frac{1}{n_s}\right) - \frac{1}{N}\right]S_y^2\right]$$

$$= \left(1 - \mathrm{P}(A1)\right)P_y^2 + \mathrm{P}(A1)\left[\mathrm{E}_1\left(\frac{1}{n_s}\right) - \frac{1}{N}\right]S_y^2$$

Calculons finalement la variance non conditionnelle. On sait que

$$Var\left(p_{ys}\right) = EQM\left(p_{ys}\right) - \left[Biais\left(p_{ys}\right)\right]^2$$

$$= \left(1 - \mathrm{P}(A1)\right)P_y^2 + \mathrm{P}(A1)\left[\mathrm{E}_1\left(\frac{1}{n_s}\right) - \frac{1}{N}\right]S_y^2 - \left[P_y - P(A1)P_y\right]^2$$

$$= \left(1 - \mathrm{P}(A1)\right)P_y^2 + \mathrm{P}(A1)\left[\mathrm{E}_1\left(\frac{1}{n_s}\right) - \frac{1}{N}\right]S_y^2 - P_y^2\left(1 - P(A1)\right)^2$$

$$= P_y^2\left(\mathrm{P}\left(A1^c\right) - \mathrm{P}\left(A1^c\right)^2\right) + \mathrm{P}(A1)S_y^2\left[\mathrm{E}_1\left(\frac{1}{n_s}\right) - \frac{1}{N}\right]$$

$$= \mathrm{P}\left(A1^c\right)P_y^2\left(1 - \mathrm{P}\left(A1^c\right)\right) + \mathrm{P}(A1)S_y^2\left[\mathrm{E}_1\left(\frac{1}{n_s}\right) - \frac{1}{N}\right]$$

$$= \mathrm{P}(A1)\left[\mathrm{P}\left(A1^c\right)P_y^2 + S_y^2\left[\mathrm{E}_1\left(\frac{1}{n_s}\right) - \frac{1}{N}\right]\right]$$

$$= \mathrm{P}(A1)\left(1 - \mathrm{P}(A1)\right)P_y^2 + \mathrm{P}(A1)\left[\mathrm{E}_1\left(\frac{1}{n_s}\right) - \frac{1}{N}\right]S_y^2$$

■

Il est à noter que cette démonstration est valide quelle que soit la distribution de probabilité de n_S. Or,

$$E_1\left(\frac{1}{n_s}\right) = \sum_{j=1}^{N}\left(\frac{1}{j}\right)\binom{N}{j}\frac{p^j(1-p)^{N-j}}{P(A1)}.$$

On remarque cependant que la quantité $E_1\left(\frac{1}{n_s}\right)$ peut rapidement devenir fastidieuse à évaluer. Il nous faut donc trouver un moyen de contourner cette difficulté. Nous allons approximer cette quantité à l'aide d'un développement en série de Taylor.

THÉORÈME 7.2.4

Peu importe la loi de n_S, si on a $0 < n_s < 2E[n_s]$, on peut faire l'approximation suivante :

$$E\left(\frac{1}{n_s}\right) \cong \left(\frac{1}{E[n_s]}\right)\left(1+\left(\frac{1}{E[n_s]}\right)^2 Var[n_s]\right)$$

Démonstration :

Dans un premier temps, développons le terme $\left(\frac{1}{n_s}\right)$ de la façon suivante :

$$\left(\frac{1}{n_s}\right) = \left(\frac{1}{E[n_s]}\right)\left(\frac{1}{1+\frac{n_s - E[n_s]}{E[n_s]}}\right) \quad \begin{cases} \text{Posons } x = \dfrac{n_s - E[n_s]}{E[n_s]}, \\ \text{on obtient alors} \end{cases}$$

$$= \left(\frac{1}{E[n_s]}\right)(1+x)^{-1} \quad \begin{cases} \text{Mais selon Taylor on peut écrire que} \\ (1+x)^{-1} = \left(1 - x + x^2 - x^3 + \ldots\right), \text{ d' où} \end{cases}$$

$$\left(\frac{1}{n_s}\right)=\left(\frac{1}{E\left[n_s\right]}\right)\left(1-x+x^2-x^3+...\right)$$

Ainsi, $\frac{1}{n_s}\cong\left(\frac{1}{E\left[n_s\right]}\right)\left(1-x+x^2\right)$ si l'on néglige tous les termes d'ordre supérieur, pourvu que $|x|<1$, c'est à dire $0<n_s<2E\left[n_s\right]$. Cette dernière condition est généralement utilisée par plusieurs auteurs en particulier SÄRNDAL [7].

$$\approx\left(\frac{1}{E\left[n_s\right]}\right)\left(1-x+x^2\right)$$

$$=\left(\frac{1}{E\left[n_s\right]}\right)\left(1-\frac{n_s-E\left[n_s\right]}{E\left[n_s\right]}+\left(\frac{n_s-E\left[n_s\right]}{E\left[n_s\right]}\right)^2\right)$$

Nous avons alors l'approximation d'ordre 2 suivante de $\mathrm{E}\left(\frac{1}{n_s}\right)$. En effet,

$$E\left(\frac{1}{n_s}\right)\cong E\left[\left(\frac{1}{E\left[n_s\right]}\right)\left(1-\frac{n_s-E\left[n_s\right]}{E\left[n_s\right]}+\left(\frac{n_s-E\left[n_s\right]}{E\left[n_s\right]}\right)^2\right)\right]$$

$$=\left(\frac{1}{E\left[n_s\right]}\right)\left(1-\left(\frac{1}{E\left[n_s\right]}\right)E\left[n_s-E\left[n_s\right]\right]+\left(\frac{1}{E\left[n_s\right]}\right)^2E\left[\left(n_s-E\left[n_s\right]\right)^2\right]\right)$$

$$=\left(\frac{1}{E\left[n_s\right]}\right)\left(1+\left(\frac{1}{E\left[n_s\right]}\right)^2Var\left[n_s\right]\right)$$

■

En particulier, pour le plan aléatoire de Bernoulli où n_s suit une loi binomiale non tronquée telle que $E[n_s]=Np$ et $Var(n_s)=Np(1-p)$, on obtient que

$$\mathrm{E}\left(\frac{1}{n_s}\right)\cong\frac{1}{n}+\frac{1-p}{n^2}\quad .$$

En supposant que $P(A1) \cong 1$, nous obtenons une approximation du second degré de la variance non conditionnelle :

$$Var(p_{ys}) \cong \frac{1-p}{n}\left[1+\frac{1}{n}\right]S_y^2.$$

7.2.2 Le plan post-stratifié

Comme nous l'avons déjà mentionné, l'estimateur de la proportion que propose le plan de post-stratification est le même que celui de la stratification:

$$p_{ys.pstr} = \frac{1}{N}\sum_{h=1}^{L} N_h p_{yhs}$$

où $p_{yhs} = \frac{1}{n_{hs}}\sum_{i=1}^{n_{hs}} y_{is}$. Cependant, on remarque que la taille de l'échantillon de la strate h, n_{hs}, est aléatoire. En effet, n_{hs} représente le nombre d'unités qui se retrouvent après coup dans la strate h ($1 \le h \le L$). Pour se persuader que n_{hs} est bien une variable aléatoire, il suffit de remarquer que cette quantité variera d'un échantillon à l'autre. La variance théorique que propose la plan de stratification contient donc un terme aléatoire et il est alors impossible d'élaborer un intervalle de confiance à l'aide de cette variance. Telle est la problématique. Le plan de Bernoulli va nous permettre de pallier ce problème.

Rappelons que dans le plan de Bernoulli, la taille de l'échantillon, n_s, suivait une loi binomiale de paramètres N et $p = {E(n_s)}/{N}$. L'espérance et la variance non conditionnelles pour l'estimateur de la proportion étaient respectivement

$$E\left(p_{ys}\right) = P(A1)P_{ys} \text{ et } Var\left(p_{yhs}\right) = P(A1)\left(1-P(A1)\right)P_{yh}^2 + P(A1)\left[E_1\left(\frac{1}{n_{hs}}\right)-\frac{1}{N}\right]S_{yh}^2.$$

Dans le plan de post-stratification, nous prenons un échantillon de taille n, connue, dans la population. Ensuite, nous effectuons les groupements que nous appelons strates. De là, on remarque que la taille, n_{hs}, de chacune des strates est aléatoire. Il s'agit d'un plan de taille aléatoire où la variable n_{hs} suit une loi hypergéométrique de paramètres $N, \frac{N_h}{N} = W_h, n$ que nous

notons $H(N,W_h,n)$. Plus précisément, n_{hs} est le nombre d'objets possédant une propriété prédéfinie obtenu lors d'un tirage sans remise de n objets parmi N dont une proportion W_h d'objets possèdent la propriété en question. On note alors que $E[n_{hs}]=nW_h$ et que

$$Var(n_{hs})=nW_h(1-W_h)\left(\frac{N-n}{N-1}\right).$$

Dans le cas de la post-stratification, on remarque qu'une strate n'existe que si sa taille est supérieure ou égale à 1. Comme pour le plan de Bernoulli, définissons A1 comme étant l'événement $n_{hs}\geq 1$. Dans le plan post-stratifié, P(A1) =1 a posteriori. Si nous adaptons les équations précédentes à notre plan d'échantillonnage post-stratifié, nous obtenons que l'espérance et la variance non conditionnelles dans une strate peuvent s'écrire de la façon suivante :

$$\mathrm{E}(p_{yhs.pstr})=P_{yh} \text{ et } Var(p_{yhs.pstr})=\left[\mathrm{E}\left(\frac{1}{n_{hs}}\right)-\frac{1}{N_h}\right]S_{yh}^2.$$

Comme les populations des strates sont indépendantes les unes des autres et qu'elles proviennent d'une même population, la variance pour l'estimateur de la proportion est la somme des variances de chacune des strates :

$$Var(p_{ys.pstr})=\sum_{h=1}^{L}W_h^2\left[\mathrm{E}\left(\frac{1}{n_{hs}}\right)-\frac{1}{N}\right]S_{yh}^2$$

Dans cette expression, il ne nous reste qu'à estimer $\mathrm{E}\left(\frac{1}{n_{hs}}\right)$ où nous avons bien sûr $E[n_{hs}]=n$.

Rappelons que nous avons réussi précédemment à approximer cette dernière quantité par

$$E\left(\frac{1}{n_{hs}}\right)\cong\left(\frac{1}{E[n_{hs}]}\right)\left(1+\left(\frac{1}{E[n_{hs}]}\right)^2 Var[n_{hs}]\right)=\left(\frac{1}{nW_h}\right)\left(1+\left(\frac{1}{nW_h}\right)^2 nW_h(1-W_h)\left(\frac{N-n}{N-1}\right)\right).$$

THÉORÈME 7.2.6

Une approximation de la variance théorique de l'estimateur de la proportion pour le plan de post-stratifié est donné par :

$$Var\left(p_{ys.pstr}\right) \cong \sum_{h=1}^{L} W_h S_{yh}^2 \left(\frac{1}{n}\right) - \sum_{h=1}^{L} W_h^2 S_{yh}^2 \frac{1}{N_h} + \sum_{h=1}^{L} S_{yh}^2 \left(\frac{1}{n}\right)^2 \left(1 - W_h\right)$$

Démonstration :

$$Var\left(p_{ys.pstr}\right) = \sum_{h=1}^{L} W_h^2 \left[E\left(\frac{1}{n_{hs}}\right) - \frac{1}{N_h} \right] S_{yh}^2$$

Mais on sait que

$$E\left(\frac{1}{n_{hs}}\right) \approx \left(\frac{1}{nW_h}\right)\left(1 + \left(\frac{1}{nW_h}\right)^2 nW_h\left(1 - W_h\right)\left(\frac{N-n}{N-1}\right)\right)$$

$$\cong \sum_{h=1}^{L} W_h^2 S_{yh}^2 \left(\frac{1}{nW_h}\right)\left(1 + \left(\frac{1}{nW_h}\right)^2 nW_h\left(1 - W_h\right)\left(\frac{N-n}{N-1}\right)\right) - \sum_{h=1}^{L} W_h^2 S_{yh}^2 \frac{1}{N_h}$$

$$\cong \sum_{h=1}^{L} W_h^2 S_{yh}^2 \left(\frac{1}{nW_h}\right) + \sum_{h=1}^{L} W_h^2 S_{yh}^2 \left(\frac{1}{nW_h}\right)^2 \left(1 - W_h\right)\left(\frac{N-n}{N-1}\right) - \sum_{h=1}^{L} W_h^2 S_{yh}^2 \frac{1}{N_h}$$

$$\cong \sum_{h=1}^{L} W_h S_{yh}^2 \left(\frac{1}{n}\right) - \sum_{h=1}^{L} W_h^2 S_{yh}^2 \frac{1}{N_h} + \sum_{h=1}^{L} S_{yh}^2 \left(\frac{1}{n}\right)^2 \left(1 - W_h\right)\left(\frac{N-n}{N-1}\right)$$

Mais lorsque N est grand, il se trouve que $\left(\frac{N-n}{N-1}\right) \cong 1$

$$\cong \sum_{h=1}^{L} W_h S_{yh}^2 \left(\frac{1}{n}\right) - \sum_{h=1}^{L} W_h^2 S_{yh}^2 \frac{1}{N_h} + \sum_{h=1}^{L} S_{yh}^2 \left(\frac{1}{n}\right)^2 \left(1 - W_h\right)$$

■

On remarque que cette expression de la variance met davantage en relief les catégories où peu d'individus se retrouvent. L'expression de cette variance possède en quelque sorte une certaine vertu d'honnêteté qui lui donne un avantage sur les calculs qui négligent le fait que la taille de l'échantillon est aléatoire.

Connaissant la variance théorique de notre estimateur de la proportion, nous obtenons finalement l'intervalle de confiance en utilisant la forme suivante :

$$p_{ys,pstr} \pm z_{\alpha/2} \sqrt{\hat{V}ar\left(p_{ys,pstr}\right)},$$

où l'estimateur de la variance théorique est donné par :

$$\hat{V}ar\left(p_{ys.pstr}\right) \equiv \left(\frac{N-n}{n}\right) \sum_{h=1}^{L} s_{yh}^2 \left(\frac{1}{N_h}\right) + \frac{N^2}{n^2} \sum_{h=1}^{L} \frac{s_{yh}^2}{N_h^2} \left(1 - W_h\right)$$

et où $s_{yh}^2 = \frac{n_{hs}}{n_{hs} - 1}\left[p_{yhs}\left(1 - p_{yhs}\right)\right]$.

Conclusion

Après avoir présenté l'importance de bien établir les objectifs d'une étude, nous nous sommes arrêtés à établir une bonne définition de la population. Nous avons ensuite regardé les formes générales des bases de sondages, ce qui nous a amené à comprendre deux choses : d'une part il est impossible de trouver une base de sondage parfaite, il nous faut donc nous contenter de la base la moins imparfaite, d'autre part l'information auxiliaire pertinente peut nous aider grandement à analyser la population sous étude. Après, nous avons abordé les plans d'échantillonnage et nous avons justifié notre choix en fonction des objectifs de l'étude et de l'information auxiliaire disponible. Par la suite, nous avons calculé la taille de l'échantillon en fonction de la précision et des budgets disponibles. Sur la question de la précision, nous concluons que pour améliorer la qualité des résultats d'une enquête il faut nécessairement passer soit par la collecte d'information auxiliaire qui nécessite des coûts, soit par l'augmentation de la taille de l'échantillon. Ainsi, la théorie de l'échantillonnage se range au vieux proverbe qui mentionne "On n'a rien sans rien".

Le chapitre consacré à la conception du questionnaire, nous a amené à considérer deux points bien distincts. D'une part, nous avons remarqué qu'il existe plusieurs façons de classer les questions. La pratique courante est de les classer en fonction de l'interprétation que le responsable de l'enquête aura à faire des réponses que le répondant aura fournies. Plus précisément, il y a deux grandes catégories de questions : les questions dont les réponses doivent être interprétées de façon objective et celles dont les réponses doivent être interprétées de façon subjective. D'autre part, nous nous sommes arrêtés sur les difficultés entourant la formulation des questions et la préparation du questionnaire. De façon générale, le statisticien cherchera à produire un questionnaire répondant autant que possible aux caractéristiques suivantes : les questions sont courtes, claires et précises; le niveau de difficulté des questions est bas, le temps de remplissage du questionnaire est court, si possible, moins de dix minutes. En somme, l'étape de la préparation des questions et du questionnaire n'est pas une tâche facile mais les efforts qu'elle exige seront largement compensés par la qualité des résultats obtenus.

Le chapitre sur le traitement des données nous a amené à explorer essentiellement trois directions. Premièrement, nous avons souligné l'importance d'utiliser un mode de collecte adapté à l'étude en cours. En effet, ce qui marche avec un mode peut s'avérer catastrophique avec un autre. Deuxièmement, nous avons ensuite compris que pour diminuer l'erreur d'échantillonnage associée aux erreurs d'observations, la collecte et la codification des données doivent être suivies par une phase de contrôle. Finalement, nous avons remarqué que la non-réponse introduit systématiquement un biais et donc une diminution de précision. Il est donc important d'utiliser les bonnes méthodes pour traiter la non-réponse. En résumé, pour une enquête par sondage, l'erreur totale est la somme de l'erreur d'échantillonnage, de l'erreur d'observation, de l'erreur due à la non couverture et à la non-réponse.

À l'aide du plan de Bernoulli, nous avons mis au point de façon rigoureuse les formes d'intervalles de confiance que nous avons utilisés en pratique dans cette enquête. La justesse des approximations dépend principalement des tailles n_h réellement observées. Dans notre cas et pour chaque question ou sous question les valeurs de n_h ont permis d'obtenir des intervalles de confiance pratiques fiables.

Finalement, en connaissant mieux l'opinion des familles habitant Rock Forest, cette étude a permis aux autorités municipales de prendre des décisions conséquentes. Mentionnons que toute la force et la crédibilité qu'on peut associer à cette étude est tout à l'avantage du sérieux qui a été investi par la municipalité et à la grande réceptivité de la population de Rock Forest.

Annexe A

Les questionnaires de l'enquête

Questionnaire de l'étude sur le transport en commun de la ville de Rock Forest

1. Est-ce que vous ou quelqu'un de votre famille prenez, ou pensez prendre l'autobus comme moyen de transport?

   - oui
   - non

2. Si non, supposons que l'autobus était plus accessible, est-ce que vous ou quelqu'un de votre famille prendriez l'autobus comme moyen de transport?

   - oui
   - non

3. Combien de personnes de votre famille prennent ou pensent prendre (prendraient) l'autobus?

   - 1
   - 2
   - 3
   - 4 et plus

4. Le membre de votre famille qui utilise, ou utilisera (utiliserait) le plus souvent le réseau de transport en commun est ou sera (serait) un passager de type

   - régulier (plus de 10 passages par semaine)
   - semi-régulier (entre 6 et 9 passages par semaine)
   - occasionnel (entre 0 et 5 passages par semaine)

5. Quels sont ou seront (seraient) les trajets-type des membres de votre famille?

dép.	dest.
dép.	dest.
dép.	dest.
dép.	dest.

6. Généralement, pour votre famille, pour quelle(s) raison(s) prenez-vous ou allez-vous prendre (prendriez-vous) l'autobus?

ne pas lire

- le travail
- les études
- le magasinage
- les loisirs
- par affaire
- autres :

7. Présentement, pour votre famille, diriez-vous que dans leur ensemble les trajets du réseau de transport en commun pour la ville de Rock Forest sont

- excellents
- très bons
- bons
- mauvais
- très mauvais
- *pas d'opinion*

8. Présentement, pour votre famille, diriez-vous que votre niveau de satisfaction, en ce qui concerne le transport en commun de la ville de Rock Forest, est

- très élevé
- élevé
- bon
- peu élevé
- très peu élevé
- *pas d'opinion*

9. À propos du transport en commun pour la ville de Rock Forest, avez-vous des suggestions concernant de nouvelles destinations, de nouveaux trajets ou de nouveaux horaires?

-
-

10. Habitez-vous cette résidence pendant l'hiver? (Question pour les riverains seulement)

- oui
- non

Questionnaire de l'étude sur la bibliothèque municipale de Rock Forest

1. Est-ce que vous ou quelqu'un de votre famille savez où se trouve la bibliothèque municipale de Rock Forest?

   - oui
   - non

2. Est-ce que vous ou quelqu'un de votre famille êtes des usagers de la bibliothèque?

   - oui
   - non

3. Si oui, combien de fois y êtes-vous allés le mois dernier?

   - 0
   - 1
   - 2
   - 3 et plus

4. Diriez-vous que la qualité des services offerts par la bibliothèque est

   - excellente
   - très bonne
   - bonne
   - mauvaise
   - très mauvaise
   - *pas d'opinion*

5. Diriez-vous que la variété ou la sélection des livres est

- très satisfaisante
- satisfaisante
- peu satisfaisante
- *pas d'opinion*

6. Que pensez-vous de l'éventualité d'un déménagement de la bibliothèque aux Terrasses Rock Forest? Êtes-vous

- tout à fait d'accord
- d'accord
- pas du tout d'accord
- *pas d'opinion*

7. Avez-vous des suggestions?

Bibliographie

[1] ARDILLY (P.), Les techniques de sondage, (Éditions Technip, Paris, 1994)

[2] COCHRAN (W.), Sampling Techniques, (Wiley, New York, 1977)

[3] DESABIE (J.), Théorie et pratique des sondages, (Dunod, Paris, 1966)

[4] HOGG (R.), TANIS (E.), Probability and Statistical Inference, (Macmillan Publishing Company, New York, 1988)

[5] MEQ, Conseils pratiques pour la construction d'un instrument de mesure (fascicule 5), (Service général des communications du MEQ, Québec, 1981)

[6] RUBIN (D.), Multiple Imputation in Nonresponse in Surveys, (Wiley, New York, 1987)

[7] SÄRNDAL (C.) et al., Model Assisted Survey Sampling, (Springer-Verlag, New York, 1992)